Research Article

Identification of Key Impacting Sectors Driving S&P 500 Variation using Statistical Modeling

Rohan Jha[1], Rishabh Jha[2]

[1]Cinco Ranch High School, 23440 Cinco Ranch Blvd, Katy, TX 77494, USA
[2]Rodger and Ellen Beck Junior High School, 5200 S Fry Blvd, Katy, TX 77450, USA

ABSTRACT

We analyze the Standard and Poor's 500 (S&P 500) variation with all eleven S&P sectors. There are effectively nine sectors because we combined real estate with financials, and communication with technology. We performed regression analysis and found that all sectors, except utilities, are statistically significant predictors of the S&P 500 variation, with p-values less than 0.05. However, the impact of the technology sector is lower at only ~10% because its impact is strongly correlated with other sectors except energy and financial. We subsequently used only three independent variables: the technology, energy, and financials sectors. The regression analysis revealed they are statistically significant with p-values less than 10^{-10} and an R-square greater than 0.98. The technology sector covers over 50% of S&P variations, the financials sector covers ~35%, and energy comprises the remaining ~15%. These were validated by taking different frequencies such as monthly and weekly over time spans of the last 20, 15, and 8 years of data. Thus, we analyzed S&P 500 variability with key sectors like technology, financials, and energy. Due to economic, market, and technological interconnection, most sectors are related. Financials offer access to capital and energy is a significant part of the cost across sectors. The energy sector is also driven by global supply and demand dynamics, geopolitics, and OPEC (Organization of the Petroleum Exporting Countries) policies.

Keywords: S&P 500; S&P sectors; Statistical model; Collinearity; Regression

INTRODUCTION

The stock market offers an economic ecosystem where investors can trade stocks to grow their financial wealth. Publicly traded companies issue stocks. There are hundreds of such companies in the US, thereby making it difficult to assess the stock market. After purchasing stocks, people own a part of the company. Investors use indicators to see how the market is doing. One of these indicators is an index, a combination of many stocks. One of the most prominent indices is the Standard and Poor's 500 (S&P 500), which tracks the stock performance of the 500 largest companies listed on stock exchanges in the United States. The S&P 500 thus indicates how the market is doing due to the sheer number of diverse, large stocks in the index.

S&P 500 companies are divided into eleven different

Corresponding author: Rohan Jha, E-mail: rohansuman0521@gmail.com.

Received August 14, 2024**; Accepted** September 4, 2024
https://doi.org/10.70251/HYJR2348.234349

sectors. These sectors are materials, industrials, financials, energy, consumer discretionary, information technology, communication services, health care, consumer staples, utilities, and real estate. Two of them, real estate and communications, started recently in 2018. The materials sector is the business of transferring raw materials to a usable good. Examples are Sherwin-Williams and Dow Inc. The industrials sector produces capital goods such as building materials. Some companies include Boeing and 3M. The financials sector deals with money including banks and their subsidiaries. This includes Bank of America Corp. and Berkshire Hathaway. The energy sector primarily helps extract, refine, and distribute energy, particularly oil & gas. Some examples are Chevron and Exxon Mobil. Consumer discretionary is a direct consumer-to-company transaction, usually dealing with goods. Some examples include Tesla and Carnival. The information technology sector develops parts and overall products including technology. Some companies include Apple Inc. and Microsoft. Communication services provide networks for data exchange. This includes Verizon and AT&T. The healthcare sector includes pharmacies and healthcare services. Some companies include Pfizer and Johnson & Johnson. The consumer staples sector sells necessary goods or common luxuries such as laundry detergent. Some companies include Walmart and Proctor & Gamble. The utilities sector deals with necessary services which include gas and electricity. Some examples are Duke Energy Corp. and Exelon Corp. Lastly, the real estate sector deals with companies developing, managing, or investing in residential, commercial, and industrial properties. Some companies include Redfin Corp and Public Storage. Each of these sectors comprises several companies making analysis of these sectors along with the S&P 500 complicated.

Most people in the US invest in the market directly or indirectly through retirement or pension funds. Understanding variability in the S&P 500 is complicated, yet it is important for investors. Several researchers have attempted to predict the S&P 500 price using numerous models. Zhong and Hitchcock (1)combined technical indicators, fundamental financial data, and sentiment analysis from textual data to predict the S&P index and stock prices. Kim et al. (2) discussed the use of sentiment analysis on financial news and integrated these sentiments to predict the S&P 500 index. Phuoc et al. (3) used machine learning to predict the behavior of the Vietnam stock market. Mukherjee et al. (4) used deep learning to predict the stock market. Fuster and Zou (5)employed logistic regression and other machine-learning techniques to forecast S&P 500 price levels, using a combination of technical indicators and market data. Rodriguez et al. (6) used a machine learning approach to predict the S&P 500 absolute percent change. However, these analyses did not assess the impact of various sectors on the S&P 500. Sector understanding is fundamental in assessing S&P 500 variation. Thus, we are using a simple but foundational statistical model to understand the S&P 500 index. We combined real estate with financials and communication with technology. We thus consider nine sectors to analyze the S&P 500 using statistical modeling. We then find the key impacting sectors to capture S&P 500 behavior. Considering key sectors, we perform a regression analysis to evaluate the S&P 500 with the fewer number of variables possible. The paper starts with the Data and Methodology section. This is a basic overview of data and analysis. Then the Results and Discussion section shares the findings of the paper. Finally, the Conclusion section summarizes the results and provides drivers for the outcomes.

DATA AND METHODOLOGY

We gathered monthly and weekly data for S&P 500 and indices for all eleven S&P sectors for the last 20-year using Yahoo Finance and Python code. We performed the following steps to develop results.

Data quality control (QC)

We first performed quality control on input data obtained from Yahoo Finance with other sources such as Nasdaq. We used the S&P 500 as a dependent variable and all eleven sectors as independent variables. Since real estate and communication sectors have been introduced recently, we added real estate with financials and communication with technology. Thus, we had nine sectors in this analysis.

Identify significant variables

We then performed a regression analysis using the S&P 500 as the dependent variable and nine key sectors as independent variables. We ran a regression analysis to identify the statistically significant variables with p-values less than 0.05. For variables with p-values greater than 0.05, we systematically removed them one by one and eventually kept only statistically significant variables.

Reduce independent variables using collinearity

An objective of this modeling is to identify the key sectors impacting the S&P 500. We therefore leveraged the

concept of collinearity to meet this objective. Collinearity occurs because the independent variables used in building the regression model are correlated with each other.

Validate the model

To test the feasibility of the model, we ran the models with only key variables and assessed their p-values for statistical significance. We also performed a blind test by randomly hiding 20% of the data to test the prediction while using the other 80% to build a model. Eventually, the model is built with 100% data. We subsequently validated the outcome by varying the time durations (8, 15 and 20 years) and changing the frequency from monthly to weekly. To gain further confidence in the model, we added a dummy variable of random number between 0 and 1. We found its correlation is poor and the p-value is large ~0.75, indicating statistically insignificant.

Assess impact

Once we have a reliable regression equation, we assess the % contribution of the independent variable in forecasting. We first normalized the independent variables between 0 and 10. Accordingly, we then adjusted the coefficients of these variables. The scaled coefficients provide the impact of these independent variables in the variability of the dependent variable, S&P 500 in this case. This enables us to identify high-impacting variables.

RESULTS AND DISCUSSION

The S&P 500 measures how the top 500 companies with respect to market capitalization are performing. These 500 companies are further divided into eleven groups called sectors. The market has been growing for the last 20 years with some dips along the way like the 2008-2009 financial crisis and the 2020 Covid-19 pandemic. These eleven sectors have grown differently (Figure 1). It is noteworthy that two of the eleven sectors, real estate and communication, started recently and thus we combined them with their parent sectors. The technology sector is growing by over 20% annually. In contrast, the energy sector is growing by less than 10% annually.

Although the S&P 500 is composed of eleven sectors, it is important to know which sectors drive its variability based on historical data. To understand this, we performed a regression analysis with the S&P 500 index as the dependent variable and the nine sectors as independent variables using the last 20-year data. We found that all sectors, except the utilities sector, have p-values of less than 0.05. This suggested that the utility sector is not statistically significant for the S&P 500. The utilities sector is fundamentally different from other sectors (7). In a market-based competitive economy like the US, customers have options to select from offerings of different providers based on the quality of products/

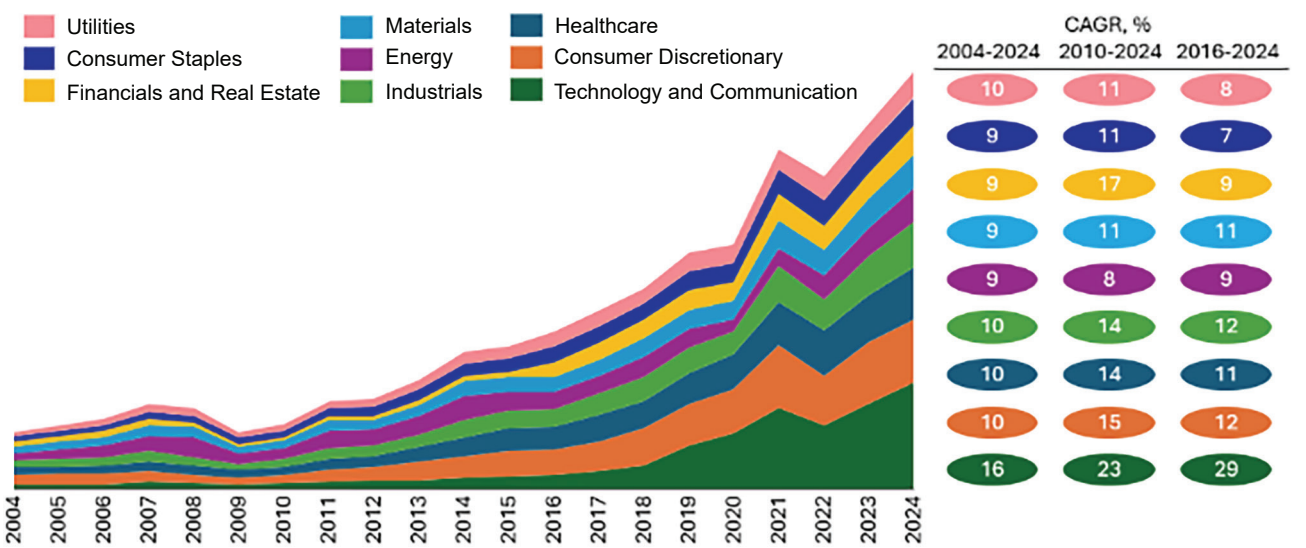

Figure 1. Index for different sectors for the last 20 years.

services, price, and other purchasing factors. Thus, supply and demand in the market set the price and if the price is too high, customers move away and vice versa. In contrast, in the US, utilities are natural monopolies since their capital costs are enormous and people cannot live without a reliable supply of utilities. To avoid the adverse impact of monopoly and the proper functioning of this sector, the US government created utility commissions to regulate the rates charged by utilities. These factors make the functioning of this sector different from that of other sectors. As a result, we then removed the utilities sector to retain only statistically significant variables. We performed regression again and found all p-values were less than 0.05. The regression output is as Multiple R as 0.998, R Square as 0.997, Adj R Square as 0.997, and Std. Error is 62.904. The regression equation is:

*S&P 500 = 457.207 + 2.150 * Technology + 3.64 * Financials + 8.52 * Healthcare + 6.177 * Consumer Discretionary - 15.074 * Consumer Staples + 2.311 * Energy + 12.007 * Industrials + 8.989 * Materials* (1)

In Equation (1), the S&P 500 is the S&P 500 US index and other sectors indices are independent variables. Subsequently, we assessed the impact of these sectors on the S&P 500. Although the regression finds the equation with all the variables, it does not mean they impact the S&P 500 equally because these indices have different ranges. To evaluate the impact of various sectors, we normalized the range of dependent variables between 0 and 10, and the normalized coefficient accordingly is used as a measure of the impact. We found that the impact of the materials sector is the lowest with less than 1% and the market cap contribution ~2% (8). The healthcare sector has the highest contribution of 29%, although its market cap contribution in the S&P 500 is ~12% (8). The highest impact of healthcare is driven by the sector having the highest overall correlation with other sectors, as revealed in the correlation Tables 2-5. Notably, technology, the biggest sector, has a relatively lower impact with only 11% on the S&P 500.

We investigated further to understand why the impact of the technology sector is lower, as shown in Table 1. This can happen if the impact of the technology sector is captured by another sector, and we could test this using collinearity. Thus, we performed collinearity among all these sectors including the S&P 500. We found that all these sectors are mostly strongly correlated (Table 2). This in turn suggests that we can run a regression with a fewer variables. We have selected three sectors: Technology, which has the biggest size, and financials and energy, which are relatively less correlated with other variables. These three values are statistically significant with p-values less than 10^{-10}. The regression output with these variables being the significant variables is R as 0.980, R square as 0.960, Adj R Square as 0.959, and Std. Error as 233.816.

The regression equation is given as:

*S&P 500 = 538.450 + 7.775 * Technology + 10.448 * Energy + 18.844 * Financials* (2)

The accuracy of the regression equation (2) is tested using a blind test by hiding ~20% of data randomly. The predicted and actual S&P 500 match reasonably well (Figure 2). The median error is ~3% and the average error is ~5%.

We then tested the collinearity using data starting from 2010 and 2016 with monthly weekly data, as shown in Tables 3-5. In each of these cases, we found that all sectors

Table 1. Impact of various sectors on S&P 500 variability using regression equation (1)

Sector	Impact on S&P 500
Technology	11%
Energy	3%
Financials	9%
Healthcare	29%
Consumer Discretionary	7%
Consumer Staples	21%
Industrials	19%
Materials	0%

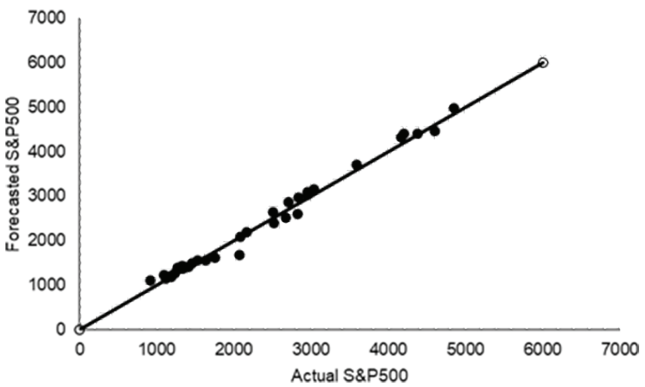

Figure 2. Actual and predicted S&P 500 are shown with bullets. The solid line is a 45-degree line.

are strongly correlated except energy. Subsequently, we took technology and financials as a measure for all other sectors and energy as dependent variables and ran the regression analysis. These regressions resulted in these three variables being significant with p-values less than 10^{-10} and R square as ~0.98. Regression equations are presented in equations (3) to (5). The regression output for data starting in 2010 and monthly frequency is Mult R as 0.977, R square as 0.956, Adj. R Square as 0.956, Std. Error as 237.7. The regression equation is:

*S&P 500 = 706.461 + 7.874 * Technology + 8.447 * Energy + 17.656 * Financials* (3)

The regression output for data starting in 2016 and monthly frequency is Mult R as 0.993, R square as

Table 2. Collinearity between different sectors for monthly data between Jan 2004 and May 2024

	S&P 500	Technology	Financials	Healthcare	Consumer Discretionary	Consumer Staples	Energy	Utilities	Industrials	Materials
S&P 500	1.00									
Technology	0.97	1.00								
Financials	0.96	0.92	1.00							
Healthcare	0.99	0.95	0.95	1.00						
Consumer Discretionary	0.99	0.96	0.96	0.99	1.00					
Consumer Staples	0.97	0.92	0.94	0.99	0.97	1.00				
Energy	0.68	0.60	0.55	0.68	0.61	0.70	1.00			
Utilities	0.97	0.92	0.95	0.98	0.97	0.99	0.67	1.00		
Industrials	0.99	0.95	0.95	0.99	0.98	0.98	0.71	0.97	1.00	
Materials	0.99	0.95	0.94	0.98	0.98	0.97	0.73	0.97	0.99	1.00

Table 3. Collinearity between different sectors for monthly data between Jan 2010 and May 2024

	S&P 500	Technology	Financials	Healthcare	Consumer Discretionary	Consumer Staples	Energy	Utilities	Industrials	Materials
S&P 500	1.00									
Technology	0.97	1.00								
Financials	0.95	0.90	1.00							
Healthcare	0.99	0.95	0.95	1.00						
Consumer Discretionary	0.99	0.97	0.95	0.98	1.00					
Consumer Staples	0.98	0.93	0.96	0.99	0.97	1.00				
Energy	0.53	0.48	0.40	0.53	0.42	0.52	1.00			
Utilities	0.96	0.92	0.97	0.98	0.96	0.99	0.47	1.00		
Industrials	0.99	0.95	0.95	0.98	0.97	0.98	0.57	0.96	1.00	
Materials	0.99	0.96	0.93	0.98	0.97	0.97	0.59	0.95	0.99	1.00

0.986, Adj. R square as 0.986, Std. Error as 105.118. The regression equation is:

$$S\&P\ 500 = 603.065 + 7.35 * Technology + 4.863 * Energy + 23.869 * Financials \quad (4)$$

The regression output for data starting in 2016 and weekly frequency is Mult R as 0.998, R square as 0.996, Adj. R square as 0.996, Std. Error as 12.405. The regression equation is:

$$S\&P\ 500 = 355.183 + 15.301 * Technology + 1.761 * Energy + 25.407 * Financials \quad (5)$$

These statistically significant variables and similar equations further provide confidence in the analysis. We

Table 4. Collinearity between different sectors for monthly data between Jan 2016 and May 2024

	S&P 500	Technology	Financials	Healthcare	Consumer Discretionary	Consumer Staples	Energy	Utilities	Industrials	Materials
S&P 500	1.00									
Technology	0.98	1.00								
Financials	0.96	0.92	1.00							
Healthcare	0.98	0.96	0.94	1.00						
Consumer Discretionary	0.97	0.97	0.94	0.93	1.00					
Consumer Staples	0.96	0.93	0.93	0.98	0.90	1.00				
Energy	0.58	0.48	0.54	0.61	0.39	0.63	1.00			
Utilities	0.90	0.90	0.92	0.94	0.87	0.96	0.54	1.00		
Industrials	0.98	0.94	0.94	0.96	0.92	0.94	0.68	0.88	1.00	
Materials	0.98	0.93	0.96	0.97	0.94	0.96	0.64	0.90	0.98	1.00

Table 5. Collinearity between different sectors for weekly data between Jan 2016 and May 2024

	S&P 500	Technology	Financials	Healthcare	Consumer Discretionary	Consumer Staples	Energy	Utilities	Industrials	Materials
S&P 500	1.00									
Technology	0.90	1.00								
Financials	(0.09)	(0.34)	1.00							
Healthcare	0.95	0.83	0.04	1.00						
Consumer Discretionary	0.98	0.95	(0.16)	0.92	1.00					
Consumer Staples	0.96	0.95	(0.28)	0.87	0.96	1.00				
Energy	0.87	0.86	(0.37)	0.86	0.87	0.88	1.00			
Utilities	0.89	0.90	(0.31)	0.91	0.89	0.90	0.95	1.00		
Industrials	0.93	0.72	0.23	0.95	0.87	0.81	0.73	0.78	1.00	
Materials	0.91	0.66	0.21	0.89	0.84	0.79	0.69	0.70	0.96	1.00

also estimated the impact of these three variables and found that it was similar to the previous cases. The technology sector covered ~50%, the financial sector covered ~35%, and the energy sector covered ~15% (Table 6). The market cap of these three sectors is ~60% of the S&P 500. The remaining sectors' market cap of ~40% is captured through these three sectors' correlation with other sectors. That is captured in the impact estimation using the regression equations. Thus, the percentage market capitalization does not correspond to percentage impact on the S&P 500.

CONCLUSIONS

In this paper, we analyzed the impact of different sectors on S&P 500 variability using statistical modeling. We found that all sectors except utilities are statistically significant for the S&P 500. Subsequently, we also found that all sectors are highly correlated except for energy and financials. Thus, we performed a regression analysis, which required only three variables such as technology and financials as a measure for other sectors and energy. These three are significant variables with p-values less than 10^{-10} and R^2 as ~0.98. The technology variable covered ~50% of variations in the S&P 500, financials covered ~35% whereas the energy sector covered the remaining ~15%. Thus, investors could focus on these key sectors to understand the S&P 500's behavior.

All sectors except for energy are strongly connected since they are coordinated with the broader economy. For example, when the economy is doing well, consumer spending and investment increase. Technological advancement infuses growth in other sectors like healthcare and financials and financials offer capital to companies and the economy. Thus, technology and financial sectors effectively represent all other sectors well. In contrast, the energy sector is driven by oil and gas prices. They are mostly driven by demand and supply dynamics, geopolitical events, and OPEC (Organization of the Petroleum Exporting Countries) policies. These factors cause the energy sector to deviate from other sectors. Particularly, energy sometime moves opposite to the economy overall. High oil and gas prices increase the production cost of all goods and services, adversely impacting them. Particularly, between 2015 to 2020, when most sectors were doing well, the energy sector was struggling due to the OPEC policy of oversupplying the market resulting in prolonged depressed oil prices.

Table 6. Impact of key sectors on S&P 500 variability and their market cap

Sector	Impact on S&P 500 using regression equations (2-5)	Market cap (8), %
Technology including communication	~50%	40%
Energy	~15%	4%
Financials including real estate	~35%	15%

ACKNOWLEDGMENTS

The authors acknowledge input and feedback from Mazhar Islam of Loyola University New Orleans, and Abhinav Jha of Stanford University. They would also like to thank their parents for their support and encouragement, and reviewers for their input to improve the quality of the paper.

REFERENCE

1. Zhong S & Hitchcock D. S&P 500 Stock Price Prediction Using Technical, Fundamental and Text Data. Statistics, Optimization & Information Computing. 2021; 9 (4): 769-788.
2. Kim J, Kim H-S, Choi S-Y. Forecasting the S&P 500 Index Using Mathematical-Based Sentiment Analysis and Deep Learning Models: A FinBERT Transformer Model and LSTM. Axioms. 2023; 12 (9): 835.
3. Phuoc T, et al. Applying machine learning algorithms to predict the stock price trend in the stock market – The case of Vietnam. Humanit Soc Sci Commun. 2024; 11, 393.
4. Mukherjee S, et al. Stock market prediction using deep learning algorithms. CAAI Transactions on Intelligence Technology. 2021; 8 (1).
5. Fuster A & Zou Z. Using Machine Learning Models to Predict S&P 500 Price Level and Spread Direction. Our code is available at: https://github.com/akfuster/CS229 (Accessed on 2024-06-25).
6. Rodriguez F, et al. A machine learning approach to predict the S&P 500 absolute percent change. Discover Artificial Intelligence. 2024; 4 (8).
7. US Utilities Sector primer. S&P Global Homepage. https://www.spglobal.com/marketintelligence/en/pages/us-utilities-sector-primer. 2022. (Accessed on 2024-09-03)
8. Sector Research. Fidelity. https://digital.fidelity.com/prgw/digital/research/sector. 2024. (Accessed on 2024-09-03)