CS982: Big Data Technologies

Report: Impact of Macroeconomic variables on The Standard and Poor's 500 (S&P 500) index

Environment – Python 3.8.8 (default, Apr 13 2021, 15:08:03) [MSC v.1916 64 bit (AMD64)]

Prepared by
Rishabh Gupta
(MSc in Quantitative Finance – 21/22)

Contents

List of figures and tables

Figures

Tables

1. Introduction

“The best thing to do is to buy 90% in an S&P 500 index fund” – This was the instruction given by Warren Buffet on 1st May, 2021 at Berkshire Hathaway’s annual meeting in Los Angeles, California to his trustee.*

These statements by legendary investors build an unbreakable trust in one of the leading economic indicators of the United states – S&P 500, which pushes me to study its backend structure.

The objective of this report is to examine the impact of macroeconomic variables on a US based stock market index – Standard and Poor’s 500, using Multiple Regression and Principal Component Analysis.

The following macroeconomic variables have been used to determine the relationship with the S&P index: -

Industrial Production index	**M1 money supply**
3-month treasury bill rate	**Moody’s AAA rating yield**
Consumer price index	**Producer price index**
3-month Certificate of deposit rate	**Consumer credit owned & securitized**
Unemployment rate	

***Taylor Locke, 2021, “3 Investing lessons Warren Buffet shared at the 2021 Berkshire Hathaway meeting”**
URL: https://www.cnbc.com/2021/05/03/investing-lessons-from-warren-buffett-at-berkshire-hathaway-meeting.html

2. Supervised learning Algorithm – Regression

Multiple regression has been used for studying the relationship between dependent and independent variables. The results of regression are interpreted on the basis of **R squared**. It explains how much variation of a dependent variable is explained by the independent variable. The higher the value, the better the model.

Terms like Null and alternate hypothesis and P value will be often used in the report. Simply speaking, Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results.

- **Null hypothesis(H_0)** says that there is no statistical significance between the two variables, which means a researcher is trying to disprove.
- An **Alternative hypothesis(H_1)** is one that states there is a statistically significant relationship between two variables.
- **P value** provides probability of the hypothesis test. P value for each independent variable tests the Null Hypothesis that there is "No Correlation" between the independent and the dependent variable. So if the P value is less than the significance level (usually 5%), the Null gets rejected, meaning there exists a relationship between the variables.
- The **significance level** is the probability of rejecting the null hypothesis when it is true.

2.1 Data description

The numbers for S&P 500 index and the explanatory variables have been taken for the period- January, 1965 to August, 2021(Frequency – Monthly)

Dependent Variable*

S&P 500 – a market-capitalization-weighted index of 500 leading publicly traded companies in the United States

Independent variables*

1. **Industrial production index (IPI)** - measures levels of production and capacity in the manufacturing, mining, electric, and gas industries, relative to a base year.
2. **Consumer price index (CPI)** - measures the average change in prices over time that consumers pay for a basket of goods and services.
3. **Producer price index (PPI)** - group of indexes that calculates and represents the average movement in selling prices from domestic production over time.
4. **M1 money supply (M1)** - refers to the total volume of money held by the public at a particular point in time in an economy.
5. **3-month treasury bill rate (3m Tbill)** - yield received for investing in a government issued treasury security that has a maturity of 3 months. It is also known as the risk free rate.
6. **3-month certificate of deposit rate (3m CD**) – When a customer opens a CD account with a bank, he/she invests a specific amount of money for a set period. The issuer pays interest at regular intervals until the date of maturity, at which time the account holder receives her original investment, plus all of the interest.
7. **Moody's AAA rating yield (AAA yield)** – Yield on corporate bonds that are rated AAA by Moody's.
8. **Consumer credit owned & securitized (Credit)** – total amount extended by banks, retailers, and others to enable consumers to purchase goods immediately and pay off the cost over time with interest.
9. **Unemployment rate** - represents the number of unemployed as a percentage of the labor force.

***Source of Definitions - Investopedia, Wikipedia, FED, FRED and bankrate.com**

2.2 Data pre-processing

As we are dealing with time series data, the problem of **spurious regression** has to be corrected first.

Spurious regression is the one where time series variables are highly correlated but are not in fact directly linked. For example - Dependent variable is "reported accidents" and Independent variable is "speeding violations". There may be a high correlation between them but what if there is a third variable that has greater influence on the reported accidents – may be Miles driven.

This is evident from the below heat map of our data:

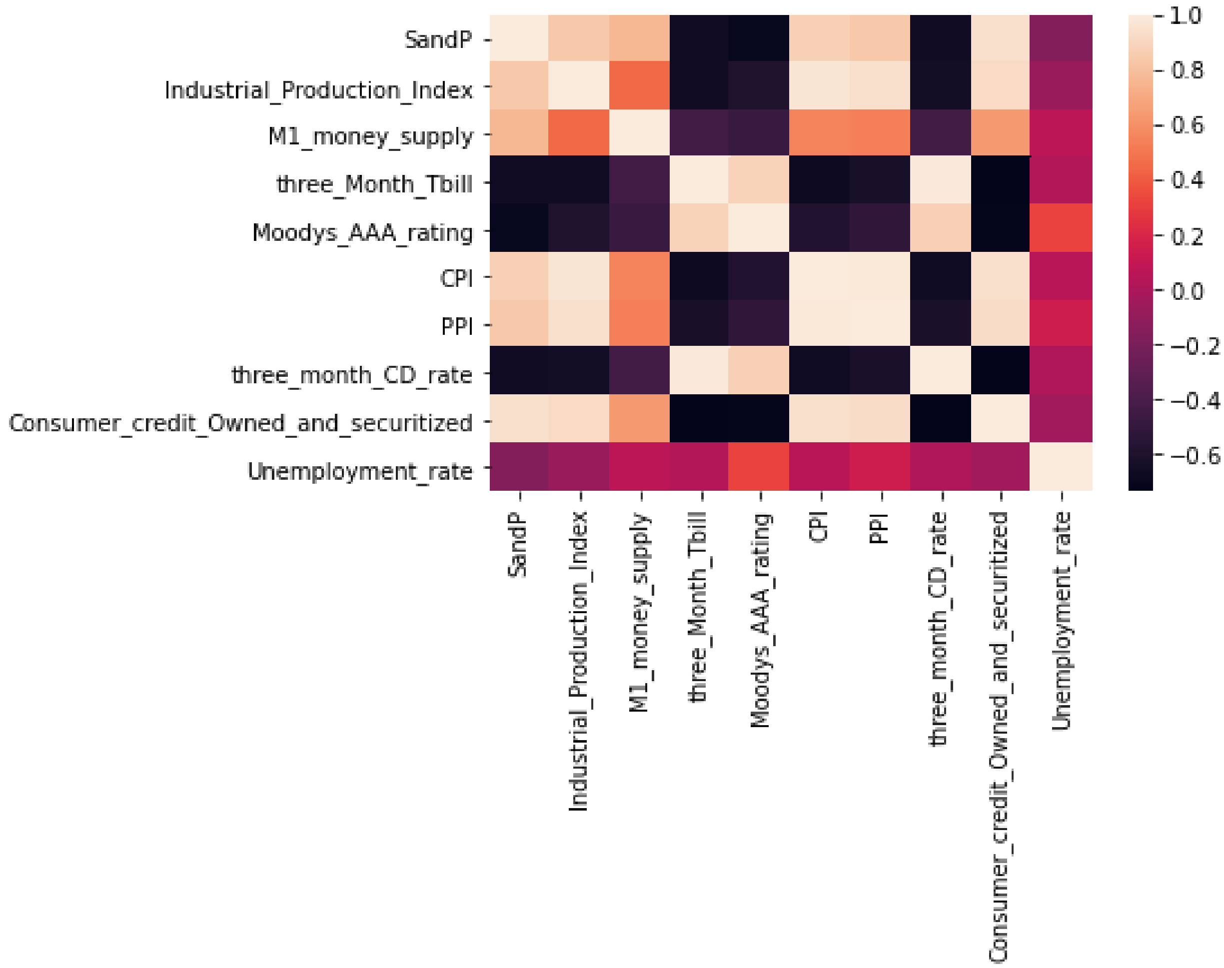

Figure 1: Correlation between all variables

We can see high correlation among CPI & PPI, SandP & CPI, SandP & PPI, 3-month t bill & 3 month CD rate and many other combinations.

2.2.1 Stationarity check and data conversion

As it is clearly evident from the plots (Figure 2), that our data has trends, we can go directly for first order differencing for converting non stationary data to stationary, instead of applying Dickey Fuller test on the raw data.

First order differencing simply means that each number is replaced by the difference between its current value and the previous value. Let's have a look at plots of some of the variables before differencing and after differencing:

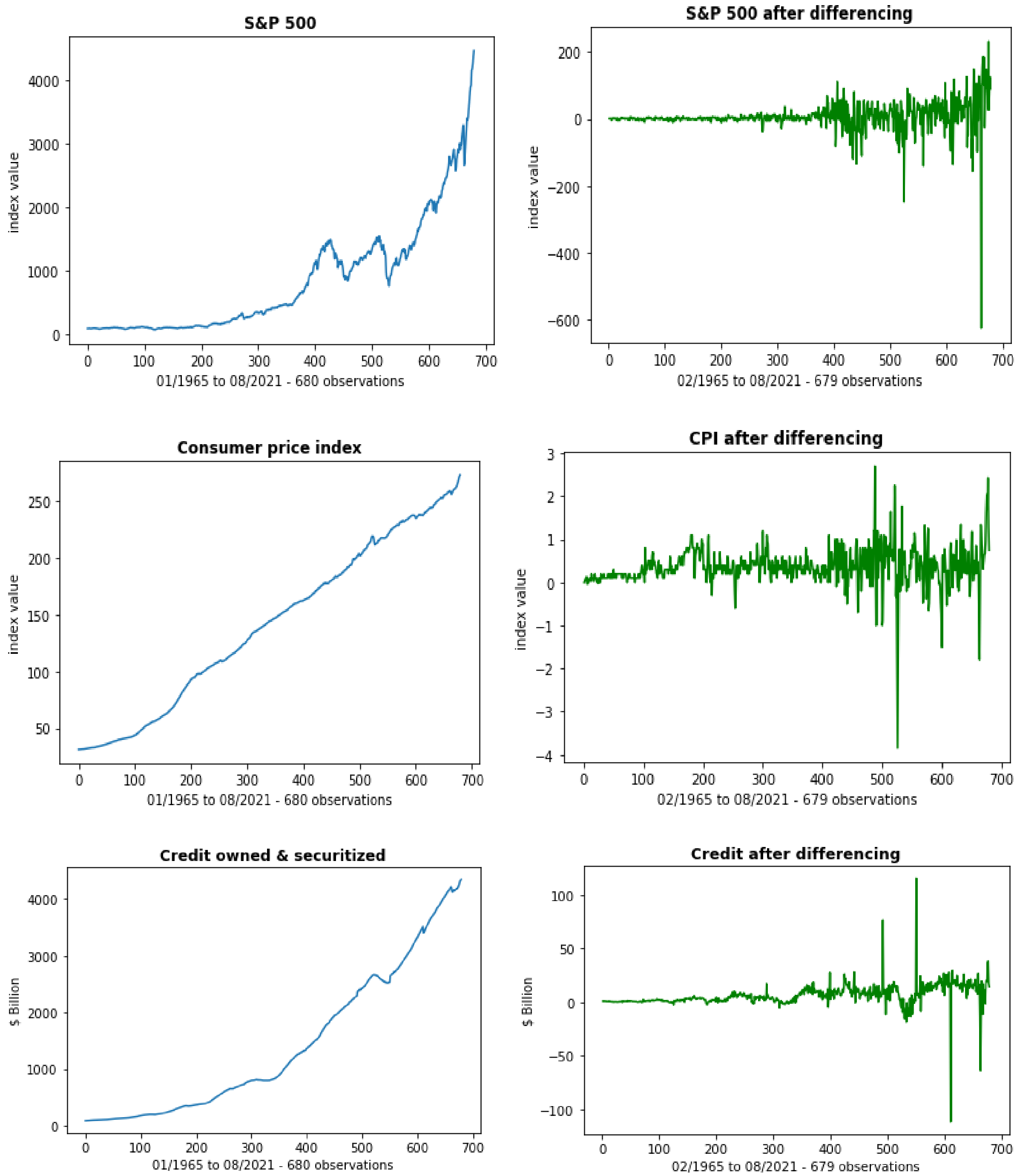

Figure 2: Glimpse of First order differencing

Since it is clearly visible from the plots that there is no trend left in the data after first order differencing, we'll still apply ADF test on this data (after differencing) to check whether there is any pattern left or not.

2.2.2 Augmented Dickey Fuller test

It is a common statistical test used **to determine whether a given time series is stationary or not**.

- **Null Hypothesis**: If failed to be rejected (p-value > 0.05), it suggests the time series is non-stationary
- **Alternate Hypothesis**: The null hypothesis is rejected (p-value <= 0.05); it suggests the time series is stationary.

After running the ADF test on the data (after First order differencing), we get p value of SandP more than 5 percent and less than 5 percent for remaining variables. It means that **S&P 500 still has some time dependent structure which needs to be removed through Second order differencing.**

Second order differencing means the same step as First order differencing and we again replace the current number by the difference between its current value and the previous value. Finally, we get the p value of SandP less than 5 percent (after second order differencing of the whole data and ADF test of SandP) and now we are all set to run the regression.

2.2.3 Outlier removal

As presented in the below boxplots, the outliers from S&P data (after second order differencing) have been removed so as to get accurate results.

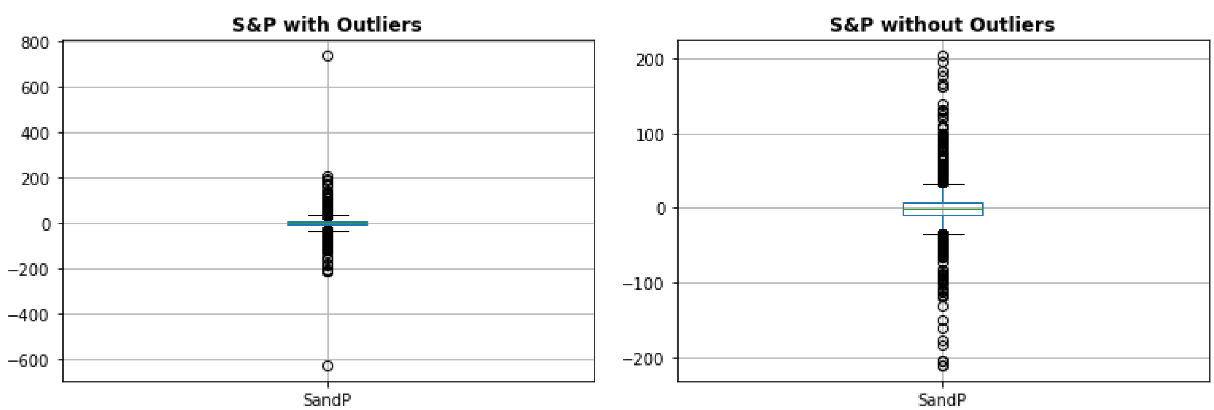

Figure 3: Effect of Outlier removal on S&P

2.3 Initial Regression Output

Our OLS equation **(Equation 1)** is as follows:

$$S\&P = \alpha + \beta_1(IPI) + + \beta_2(CPI) + \beta_3(PPI) + \beta_4(M1) + \beta_5(3m\ Tbill) + \beta_6(3m\ CD) + \beta_7(AAA\ yield) + \beta_8(Credit) + \beta_9(Unemployment\ rate)$$

Output summary:

Table 1 – Initial results

R square	**0.064**	
	Coefficient	**P value**
Intercept	-0.1639	0.923
IPI	-11.2568	0
M1	-0.0048	0.177
3m Tbill	17.809	0.002
AAA yield	-13.9564	0.061
CPI	5.7818	0.240
PPI	0.0142	0.993
3m CD	-12.2393	0.011
Credit	0.0158	0.916
Unemployment rate	-21.9026	0

- The above table shows a direct significant relationship of **CPI** with the S&P index due to high positive coefficient. The relationship of CPI can be both positive or negative with the index as it indicates inflation/deflation. So, from the above results, we can say that in the long term, when the inflation increases, the index rises.

- Also as per Rose (1994, Money and Capital Markets), "T-bill rates *typically* rise during periods of business expansion and fall during recessions", so as per this theory and my understanding, during business expansion, index will definitely rise, which exactly goes with our positive coefficient of **3m Tbill**. On the contrary, the negative coefficient of **3m CD rate** does not go with the above theory as it is also considered a safe investment.

- As the **Industrial production index** is an economic indicator measuring the real output of many major industries, it is usually expected to have a direct relation with the index, but the results above don't meet our expectations. Same is the case with **M1 money**

supply. The negative coefficient does not go with the expectation that increase in money supply increases the liquidity in the economy and encourages lending and investment and hence the index rises.

- Regarding **AAA yield**, there is nothing that can be said for/against the above result, because during economic expansion, bond yields and stock market move in the same direction whereas after a recession, bond prices and market tend to move together (meaning the yield and market moves in opposite direction). Regarding **Unemployment rate**, if we put it simply (i.e. without going deep into lowering/rising of interest rates by FED due to unemployment), the negative coefficient matches with the general expectation that low unemployment indicates economic expansion and vice versa.

- For **PPI** and **Credit,** the coefficients are almost NIL, so there is nothing to comment on.

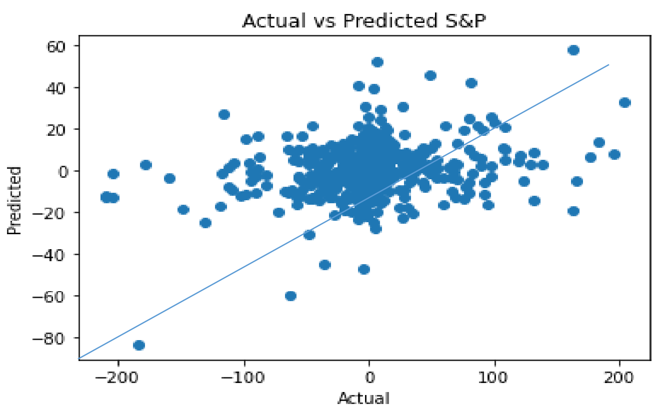

Figure 4: Initial Plot of Y and Predicted Y

The plot of Actual and Predicted values of S&P in Figure 4 is not showing a good linear relationship between the Dependent and independent variables.

A good model should show the points in the plot having scattered around a 45-degree line in a symmetrical manner.

For that, we'll have to fulfill all the assumptions of Multiple regression.

Assumptions in Regression that should not be violated:

1. There should be a linear relationship between dependent variable and independent variable(s). A linear relationship suggests that a change in response Y due to one unit change in X is constant, regardless of the value of X.
2. There should be no correlation between the error terms. Absence of this phenomenon is known as Autocorrelation.
3. The independent variables should not be correlated. Absence of this phenomenon is known as Multicollinearity.
4. The error terms must have constant variance. This phenomenon is known as homoscedasticity. The presence of non-constant variance is called heteroscedasticity.
5. The error terms must be normally distributed.

2.4 Important variable selection

Before heading to fulfilling the above assumptions, the variables that are most important for providing best fit for the model have to be selected.

The Backward Selection method has been used for removing irrelevant variables. Variables with P value more than 5 percent will be removed, as it implies that we fail to reject the Null hypothesis - meaning there does not exist a relationship between that variable and our dependent variable.

Now, from Table 1, **PPI** has the highest p value, so we'll be dropping PPI from the equation and rerun regression on the below remaining equation **(Equation 2)**:

S&P = α + β_1(IPI) + + β_2(CPI) + β_3(M1) + β_4(3m Tbill) + β_5(3m CD) + β_6(AAA yield) + β_7(Credit) + β_8(Unemployment rate)

Now, the highest p value from the results of this equation will be seen and then that variable will be removed. Now the highest p value belongs to **Credit**. Credit is removed and we'll again rerun regression on the following equation **(Equation 3)**:

S&P = α + β_1(IPI) + + β_2(CPI) + β_3(M1) + β_4(3m Tbill) + β_5(3m CD) + β_6(AAA yield) + β_7(Unemployment rate)

This is an iterative process until we reach a point where the p values of all the variables are less than 5%. Now the next candidate for removal is **M1**, then **CPI,** and then the **AAA yield.** We are left with the below significant variables after the above elimination procedure:

Table 2 – Final Significant variables

R square	**0.053**	
	Coefficient	**P value**
IPI	11.3407	0.000
3m Tbill	14.9487	0.010
3m CD	-12.8573	0.006
Unemployment rate	-18.6979	0.000

Luckily, the coefficients of the above variables go with our expectations **except 3m CD rate**.

2.5 Multicollinearity check

Multicollinearity is a problem, because we would not be able to distinguish between the individual effects of the independent variables on the dependent variable. So, instead of having two highly correlated variables, it's better to have one variable so as to protect the model from overfitting.

Variance Inflation factors (VIF) are used in this project to address this issue. VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable.

VIF = $1/(1-R^2)$

VIF more than 5 indicates high multicollinearity. We've obtained the below VIFs:

Table 3 – Variance Inflation factors

Variable	**VIF**
IPI	1.6038546778075322
3m Tbill	2.889548707788511
3m CD	2.901249810736872
Unemployment rate	1.595488052863333

Therefore, our data is free from Multicollinearity as we don't have any variable with VIF of more than 5.

2.6 Autocorrelation check

- **Durbin Watson test**

This test has been used to check correlation between the error terms. It is a statistical test based on the test statistic, which is computed by the below formula:

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2},$$

Here, e_t refers to error terms of current period and e_{t-1} means their lagged value for 1 period.

The DW test statistic varies from 0 to 4, with values between 0 and 2 indicating positive autocorrelation, 2 indicating zero autocorrelation, and values between 2 and 4 indicating negative autocorrelation. We are getting a test statistic of **2.81(after regressing with the remaining 4 independent variables), which means negative autocorrelation.**

- **Cochrane-Orcutt Procedure**

This method has been used to treat Autocorrelation. Simply speaking, it involves regressing the error term with the lagged value of the error term. Then the coefficient of the lagged error term **(r)** is used to transform our original equation. The transformed dependent and independent variables have been computed like this:

$S\&P_t^* = S\&P_t – [r\ (S\&P_{t-1})]$
$IPI_t^* = IPI_t - [r\ (IPI_{t-1})]$
$3m\ Tbill_t^* = 3m\ Tbill_t - [r\ (3m\ Tbill_{t-1})]$
$3m\ CD_t^* = 3m\ CD_t - [r\ (3m\ CD_{t-1})]$
$Unemployment\ rate_t^* = Unemployment\ rate_t - [r\ (Unemployment\ rate_{t-1})]$

Note 1 - t refers to current value and t-1 refers to lagged value for 1 period.
Note 2 – Variables with * refers to transformed variables

After running the regression on the above transformed variables, we get a DW statistic of 2.25, which is close to 2 and nearly acceptable.

The remaining process will be carried out with the below equation **(Equation 4)**, as the Autocorrelation has been almost removed.

$$\mathbf{S\&P_t^* = \alpha + \beta_1(IPI_t^*) + \beta_2(3m\ Tbill_t^*) + \beta_3(3m\ CD_t^*) + \beta_4(Unemployment\ rate_t^*)}$$

2.7 Linearity test

By looking at the below plot obtained from the predicted values of the Equation 4, we can infer that the expected value of a dependent variable is not a straight-line function of each independent variable.

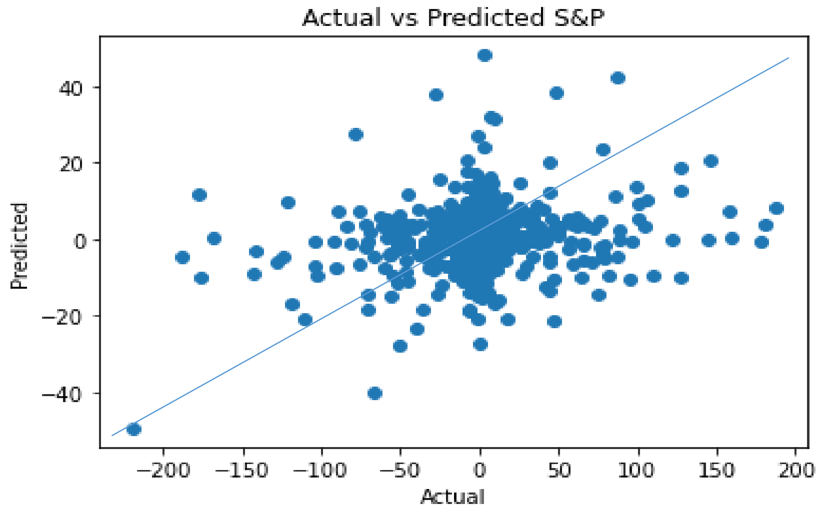

Figure 5: Final plot of Actual Y and Predicted Y

For treating Nonlinear relationships, we'll carry out the process of **Polynomial Regression**.

- **Polynomial Regression**

This process involves adding the squares of all the independent variables and then running the regression. The new equation **(Equation 5)** is as follows:

$$S\&P_t^* = \alpha + \beta_1(IPI_t^*) + \beta_2(3m\ Tbill_t^*) + \beta_3(3m\ CD_t^*) + \beta_4(Unemployment\ rate_t^*) + \beta_5(IPI_t^*)^2 + \beta_6(3m\ Tbill_t^*)^2 + \beta_7(3m\ CD_t^*)^2 + \beta_8(Unemployment\ rate_t^*)^2$$

After running the regression and selecting variables **again based on the Backward selection method**, the system has dropped all the squared variables and we are back to the same equation **(Equation 4)** and exactly the same plot as in Figure 5.

Therefore, Polynomial regression has failed to help us in removing nonlinear relationships.

2.8 Heteroscedasticity check

Breusch-Pagan Test

One way to determine if heteroscedasticity is present is to use the **Breusch-Pagan Test**. It involves using a variance function and using a Chi square test to test the following hypothesis:

- **Null** - heteroscedasticity is not present
- **Alternative** - heteroscedasticity is present

If p value is less than 5%, we reject the Null and vice versa. We are getting a p value of 0.6, which means we fail to reject the Null and our data is free from heteroscedasticity.

If we look at the **Chi square test results**, our chi square statistic (2.75) is less than chi square critical (3.84), implying no heteroscedasticity.

Normality of error terms

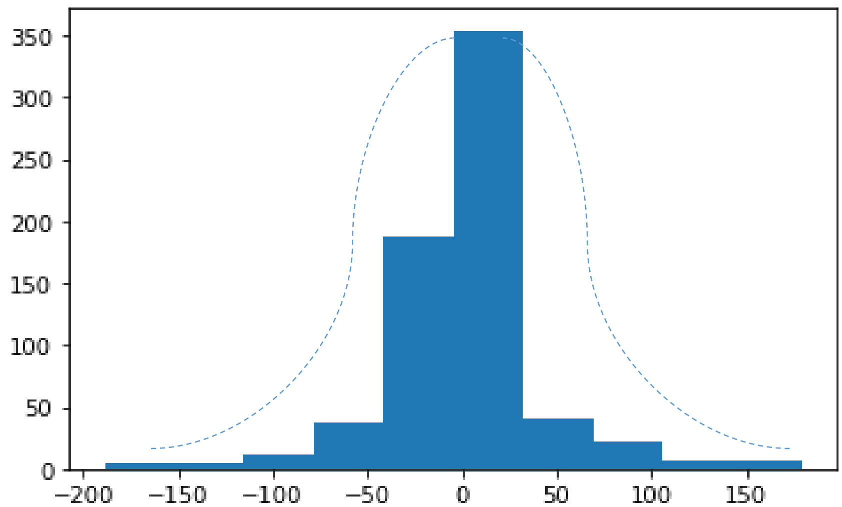

Figure 6: Error terms Histogram

Figure 6 shows the histogram plot of our residuals, which nearly seems like a Normal distribution bell curve and fulfils the assumption of achieving a normal distribution of error terms.

2.9 Conclusion

To conclude, from Equation 4, we have the following results:

Table 4 – Regression Final results

R square	**0.041**	
	Coefficient	**P value**
IPI	-9.6982	0.000
3m Tbill	17.7724	0.002
3m CD	-15.6956	0.001
Unemployment rate	-14.7124	0.001

- As per p values, all the above variables are significant for impacting the S&P index, but the coefficients are disappointing us as the coefficient for IPI and 3m CD is negative, which is against our general expectation.

- We can say that we have fulfilled the assumptions of Multicollinearity, Autocorrelation, Heteroscedasticity and normality of error terms, but the scatter plot from Figure 5 does not convey a **strict linear relation** between S&P and our four variables.

3. Principal Component Analysis (PCA) – Unsupervised learning Algorithm

PCA is specifically used to reduce dimensionality of the data consisting of a large number of interrelated variables. Here, the purpose of using PCA is just to identify the principal variables that significantly explains my target variable S&P.

After applying the algorithm on our initial dataset (without first order differencing), we get the following **Explained variance ratios** (Table 5).

Explained variance ratio is a metric to evaluate the usefulness of your principal components and to choose how many components to use in your model.

Table 5 – Explained Variance ratios

Component	**Explained variance ratio**	**Cumulative ratio**
IPI	0.666	0.666
M1	0.145	0.811
3m Tbill	0.096	0.907
AAA yield	0.074	0.981
CPI	0.009	0.99
PPI	0.004	0.994
3m CD	0.002	0.996
Credit	0.0008	0.9968
Unemployment rate	0.0005	0.9973

We are achieving 99% of the variance at CPI, so after running the final regression on our transformed data (through PCA) with the first 5 variables, the results are:

Table 6 – PCA results

R square	**0.953**	
	Coefficient	**P value**
IPI	328.62	0
M1	34.97	0
3m Tbill	-218.73	0
AAA yield	285.86	0
CPI	166.17	0

4. Results comparison

Table 7 – Final comparison

	Multiple regression	**PCA**	**Remarks**
R square	**0.041**	**0.953**	
	Coefficients:	**Coefficients:**	
IPI	-9.6982	328.62	Contradiction
M1		34.97	In line with expectation
3m Tbill	17.7724	-218.73	Contradiction
AAA yield		285.86	In line with expectation
CPI		166.17	In line with expectation
3m CD	-15.6956		Not in line with expectation
Unemployment rate	-14.7124		In line with expectation

If we filter out the variables (Table 7), it can be concluded that Money Supply, Moody's AAA rated yield, Consumer Price index and Unemployment rate most likely influence the US stock market. *

* **The aforementioned findings are only for research and analysis purposes and should not be utilised to make any financial decisions.**

5. References

Joseph Tagne Talla, 2013, Impact of Macroeconomic Variables on the Stock Market Prices of the Stockholm Stock Exchange(OMXS30)

Shengxuan Chen, 2020, Forecasting Daily Stock Market Return with Multiple Linear Regression

Stephanie Glen. "Hypothesis Testing" From StatisticsHowTo.com: Elementary Statistics for the rest of us!
URL: https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/

"What is a Null Hypothesis? - Definition & Examples." *Study.com*, 5 January 2016,
URL: study.com/academy/lesson/what-is-a-null-hypothesis-definition-examples.html

Abhigyan (2020) 'Understanding the P-value in Regression'
URL: https://medium.com/analytics-vidhya/understanding-the-p-value-in-regression-1fc2cd2568af

Webcourses@UCF, 2.1, Spurious Relationships
URL: https://webcourses.ucf.edu/courses/1169320/pages/2-dot-10-spurious-relationships

Dennis Sauert (Author), 2010, Do Macroeconomic Variables have an Effect on the US Stock Market?, Munich, GRIN Verlag, https://www.grin.com/document/158131

Otexts, 8.1 – Stationarity and differencing,
URL: https://otexts.com/fpp2/stationarity.html

Introduction to Econometrics with R, 14.7 – Non stationarity 1: Trends,
URL: https://www.econometrics-with-r.org/14.7-nit.html

Jason Brownlee, 2016, How to check if Time series data is stationary with Python
URL: https://machinelearningmastery.com/time-series-data-stationary-python/

Aniruddha Bhandari, 2020, What is Multicollinearity? Here's Everything you need to know
URL: https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/

Statistics Knowledge portal, Multicollinearity
URL: https://www.jmp.com/en_gb/statistics-knowledge-portal/what-is-multiple-regression/multicollinearity.html

Anaytics Vidhya, 2016, Going Deeper into Regression Analysis with Assumptions, Plots & Solutions
URL: https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/

Statistics Knowledge portal, Regression Model Assumptions
URL: https://www.jmp.com/en_gb/statistics-knowledge-portal/what-is-regression/simple-linear-regression-assumptions.html

Kavita Mali, 2021, Everything you need to Know about Linear Regression!
URL: https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/

PennState Eberly College of Science, Lesson 9: Data Transformations
URL: https://online.stat.psu.edu/stat501/book/export/html/956

Adam Hayes, 2021, What do Constantly Low Bond Yields Mean for the Stock Market
URL: https://www.investopedia.com/ask/answers/061715/how-can-bond-yield-influence-stock-market.asp

Timothy Sykes, 2021, US Unemployment Rate: How it Can Affect Stocks
URL: https://www.timothysykes.com/blog/us-unemployment-rate/

Robert Nau, Duke University, Regression diagnostics: testing the assumptions of linear regression
URL: https://people.duke.edu/~rnau/testing.htm

PennState Eberly College of Science, 14.3 – Testing and Remedial Measures of Autocorrelation
URL: https://online.stat.psu.edu/stat501/lesson/14/14.3

Zach, 2020, How to perform a Breusch – Pagan Test in Python
URL: https://www.statology.org/breusch-pagan-test-python/

Akash Dubey, 2018, The Mathematics Behind Principal Component Analysis
URL: https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643

Isabella Lindgren, 2020, Dealing with Highly Dimensional Data using Principal Component Analysis(PCA)
URL: https://towardsdatascience.com/dealing-with-highly-dimensional-data-using-principal-component-analysis-pca-fea1ca817fe6#:~:text=The%20explained%20variance%20ratio%20is,or%2080%25%20to%20avoid%20overfitting

Jim Frost, How to Interpret R-squared in Regression Analysis
URL: https://statisticsbyjim.com/regression/interpret-r-squared-regression/

Jason Fernando, 2021, Consumer Price Index
URL: https://www.investopedia.com/terms/c/consumerpriceindex.asp

Will Kenton, 2021, Industrial Production Index
URL: https://www.investopedia.com/terms/i/ipi.asp

Source of Data

Variable	**URL**
S&P 500 indices	https://www.multpl.com/s-p-500-historical-prices/table/by-month
IPI	https://fred.stlouisfed.org/series/INDPRO
M1	https://fred.stlouisfed.org/series/M1SL
3m Tbill	https://fred.stlouisfed.org/series/TB3MS
AAA corporate bond yield	https://fred.stlouisfed.org/series/AAA
Consumer Price Index	https://fred.stlouisfed.org/series/CPIAUCSL
Producer price index	https://fred.stlouisfed.org/series/PPIACO
Certificate of deposit rates	https://fred.stlouisfed.org/series/IR3TCD01USM156N
Unemployment rate	https://fred.stlouisfed.org/series/UNRATE
Total consumer credit owned and securitized	https://fred.stlouisfed.org/series/TOTALSL