ANLY 502

Project Report

Predicting S&P 500 Index Based on Consumer-Related Indicators

Group 1: Abdul Hai Hazari; Ashok Kumar Aalla; Fan Jiang; Léa Demri

Summary

Introduction and Data Description

Stock markets are often seen as a major indicator of a country's economic health. During year 2008 financial crisis, the stock market crashed, and the U.S. GDP shrank at a dramatic 6.2%. After an almost two year's economic meltdown, starting 2010, the U.S. economy has been experiencing the longest economic expansion ever since its existence. Its GDP has been growing at a year-to-year rate around 2%, while the unemployment rate has been hovering at a constant historical low of 4%. Naturally, we want to examine if the ordinary people have benefited from the current economic boon. In our research paper, we used nine independent variables to capture the private sector. We chose S&P 500 index as our proxy for the U.S. stock market. The research interest is to examine if the changes in variables, such as Personal Consumption Expenditure, Personal Savings Rate, have a strong relationship with the movement in S&P index price. There are a total of 478 observations. We used monthly data, starting from January of 1980 to September of the current year as the sample dataset. The dependent variable, S&P500 monthly adjusted price was extracted from Yahoo Finance, while all the other consumer related data, i.e. independent variables were obtained from the website FRED Economic Data.

The multiyear regression function we used is shown as following and our aim is to find what the values of regression coefficients β_i are.

We have applied a multiple linear regression for empirical analysis to simultaneously consider the effects of each variable. The initial empirical model is given as follows:

Model 1

$\widehat{IndexPrice(monthly)}$=*$\beta_0$ + β_1Personal Consumption Expenditure+ β_2Medicare + β_3Personal Interest Income+ β_4Personal Dividend Income+ β_5Real Disposable Personal Income+ β_6Personal Savings Rate+ β_7Personal Interest PMT+ β_8Personal Current Taxes+ β_9Social Security+ ε*

The test statistic, which feasible here is T test.

The Null hypothesis is H_0: $\beta_1 = \beta_2 = \ldots = \beta_9 = 0$

The Alternative hypothesis is H_1: β_1~ $\beta_9 \neq 0$

Empirical Hypothesis and Assumptions on the Regression Model

On the basis of a literature review, the following hypothesis has been generated.

H1, H2, …, H9: There is a significant relation between $x_1, x_2, \ldots, x_9$ and *y*.

Assumptions:

1. We assume multivariate linearity.

   The nine independent variables have a significant and observable impact on the monthly S&P 500 adjusted closing price. This relationship is assumed to be linear and subject to random error.

2. We assume variance in all predictors.
   All of the nine independent variables vary.

3. We assume multivariate normality.
   The residuals are assumed to be normally distributed.

$$\varepsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n$$

4. We assume no multicollinearity.
   The nine independent variables are not highly correlated with each other.

5. We assume that the values of the residuals $(\varepsilon_1, \dots, \varepsilon_n)$ are mutually independent. Our observations must be independent from one another.

6. We assume homoscedasticity.

   The variance of error terms is constant across the values of the independent variables. A plot of standardized residuals versus predicted values shows whether points are equally distributed across all values of the independent variables.

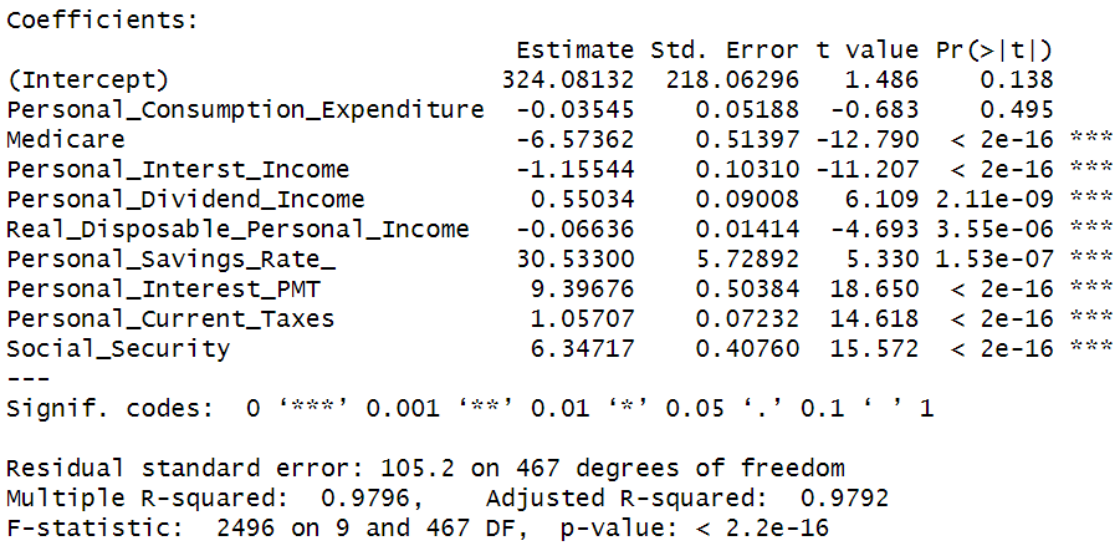
$$Var(\varepsilon_i) = \sigma^2, i = 1, \dots, n$$

7. We assume there are no influential cases biasing our model. Significant outliers and influential data points can place undue influence on the model, making it less representative of the data as a whole.

8. We assume the S&P 500 index is a good proxy for the stock market. However, since it is not mean and variance sufficient, the S&P 500 index may not be the best measure of stock market changes.

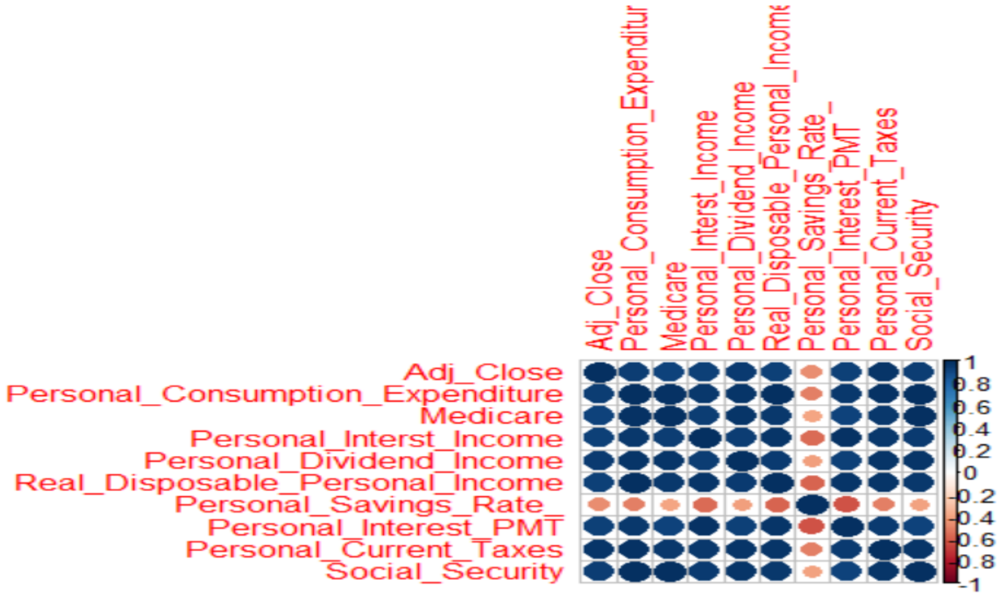
Our Findings and Explanation

As the coefficients table below shows, except the intercept and Personal Consumption Expenditure, t-stat of every coefficient, is greater than 2, therefore each 8 out of 10 individual coefficients is statistically significant. The last column, P value also proves the same result. However, because 2 out of 10 coefficients are not statistically significant, we need to drop the first variable to improve our model. The next step would be to exclude the variable Personal Consumption Expenditure, to check if this will help us to locate the best model.



```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      324.08132  218.06296   1.486    0.138    
Personal_Consumption_Expenditure  -0.03545    0.05188  -0.683    0.495    
Medicare                          -6.57362    0.51397 -12.790  < 2e-16 ***
Personal_Interst_Income           -1.15544    0.10310 -11.207  < 2e-16 ***
Personal_Dividend_Income           0.55034    0.09008   6.109 2.11e-09 ***
Real_Disposable_Personal_Income   -0.06636    0.01414  -4.693 3.55e-06 ***
Personal_Savings_Rate_            30.53300    5.72892   5.330 1.53e-07 ***
Personal_Interest_PMT              9.39676    0.50384  18.650  < 2e-16 ***
Personal_Current_Taxes             1.05707    0.07232  14.618  < 2e-16 ***
Social_Security                    6.34717    0.40760  15.572  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 105.2 on 467 degrees of freedom
Multiple R-squared:  0.9796,	Adjusted R-squared:  0.9792 
F-statistic:  2496 on 9 and 467 DF,  p-value: < 2.2e-16
```

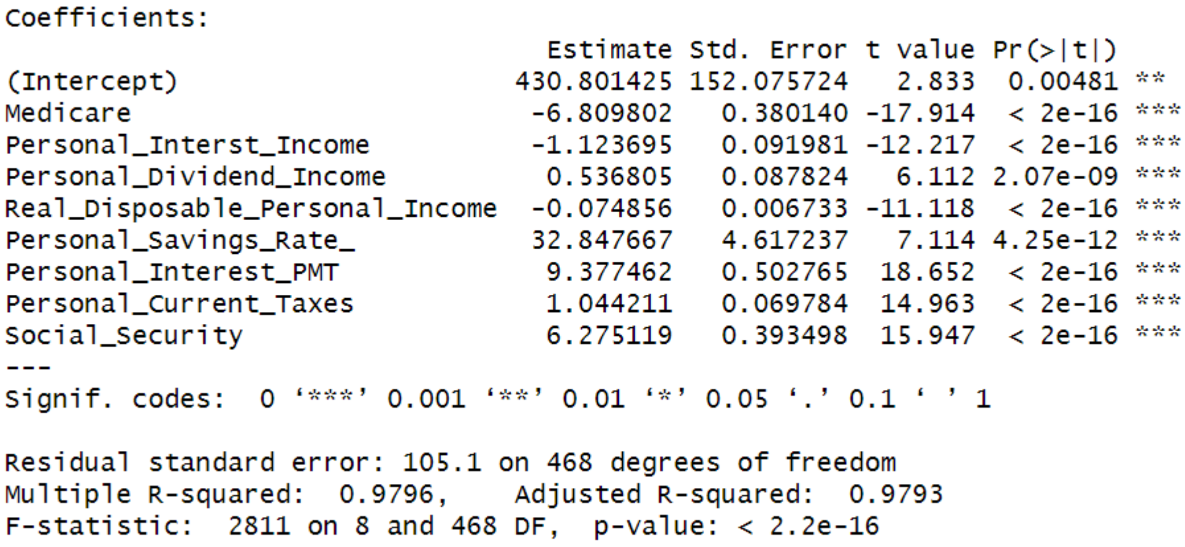


Model 2 – Best Model

$\widehat{IndexPrice(monthly)}$=$\beta_0$ + β_1*Medicare* + β_2*Personal Interest Income*+ β_3*Personal Dividend Income*+ β_4*Real Disposable Personal Income*+ β_5*Personal Savings Rate*+ β_6*Personal Interest PMT*+ β_7*Personal Current Taxes*+ β_8*Social Security*+ ε

As the coefficients table below shows, t-stat of every coefficient, is greater than 2, therefore all coefficients are statistically significant. The last column for P value also proves the same result since they all below 0.05. Therefore, we reject the Null Hypothesis which states that the coefficients are zero, and in favor of the alternative hypothesis that all our independent variables have a non-zero relation with the dependent variable.

The next step is to interpret our coefficients one by one. Starting with variable Medicare, the estimate -6.81, means an increase of 1 billion dollars in Medicare that consumers received from the government, is associate with a decrease of $6.81 in the S&P 500 index price, while holding other variables constant. The other two variables, Personal interest income and Real disposable personal income also exhibit the same negative trends. A 1 billion-dollar increase in personal interest income is associated with a drop of 1.12 dollar in the index price, while holding other variables constant. The variable that has the most significant effect on our dependent variable is the personal savings rate, which has a coefficient of 32, meaning a 1% increase in personal savings rate is associated with a jump in the index price of 32 dollars, while holding other variables constant. Another variable that also exhibit a strong relation to the index price is personal interest payment. A 1 billion-dollar increase in personal interest payment corresponds to a positive change of 9 dollars in the index. Last but not the least, it is worth to mention the coefficient for the intercept, which is 430. It indicates that without the effect of X variables in our research, the average monthly index price is at $430. However, this number will not hold out of the context. Both the Multiple R-squared and Adjusted R are at a high level of 0.97. The high score of Adjusted R^2 means that the model we use here explains the 97% of variability of the researched variable.



```
Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  430.801425 152.075724   2.833  0.00481 ** 
Medicare                      -6.809802   0.380140 -17.914  < 2e-16 ***
Personal_Interst_Income       -1.123695   0.091981 -12.217  < 2e-16 ***
Personal_Dividend_Income       0.536805   0.087824   6.112 2.07e-09 ***
Real_Disposable_Personal_Income -0.074856   0.006733 -11.118  < 2e-16 ***
Personal_Savings_Rate_        32.847667   4.617237   7.114 4.25e-12 ***
Personal_Interest_PMT          9.377462   0.502765  18.652  < 2e-16 ***
Personal_Current_Taxes         1.044211   0.069784  14.963  < 2e-16 ***
Social_Security                6.275119   0.393498  15.947  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 105.1 on 468 degrees of freedom
Multiple R-squared:  0.9796,	Adjusted R-squared:  0.9793 
F-statistic:  2811 on 8 and 468 DF,  p-value: < 2.2e-16
```

Based on the estimates of coefficients, our prediction model is:

$\widehat{IndexPrice(monthly)}$=$430 - 6.8*Medicare - 1.12*Personal Interest Income + 0.54*Personal Dividend Income - 0.08*Real Disposable Personal Income + 32.8*Personal Savings Rate + 9.37*Personal Interest PMT + 1.04*Personal Current Taxes + 6.27*Social Security + ε

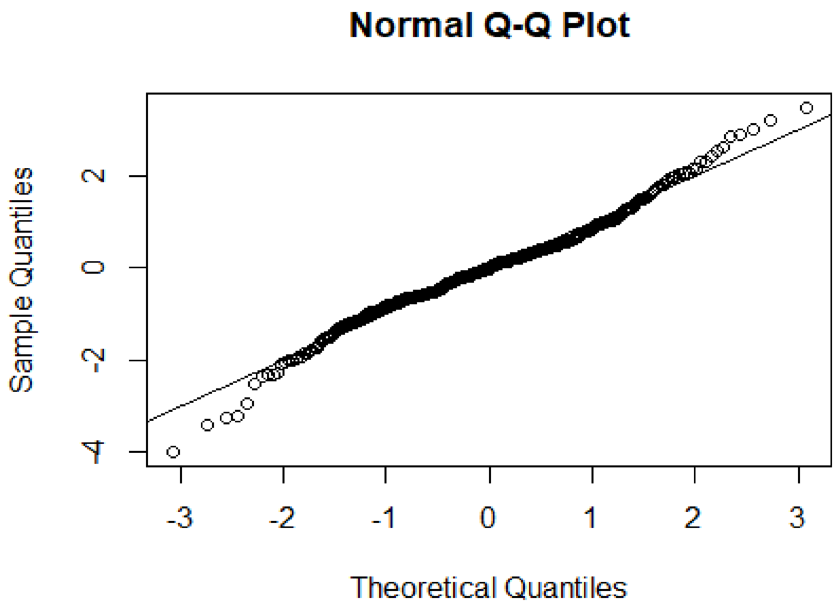
In conclusion, our model shows that the consumer-related indicators that we have selected exhibit a strong relationship with the monthly adjusted price of S &P 500. However, our research is limited by the assumption multicollinearity. We decided to include all the X variables, to keep the model as conclusive as possible, since the adjusted R is the highest with having all 8 variables in. In general, we believe the chosen X variables can be used to predict monthly movement of S&P 500.

Assumptions on the Optimal Regression Model (Model 2)

We verify the assumptions on our optimal regression model (Model 2).

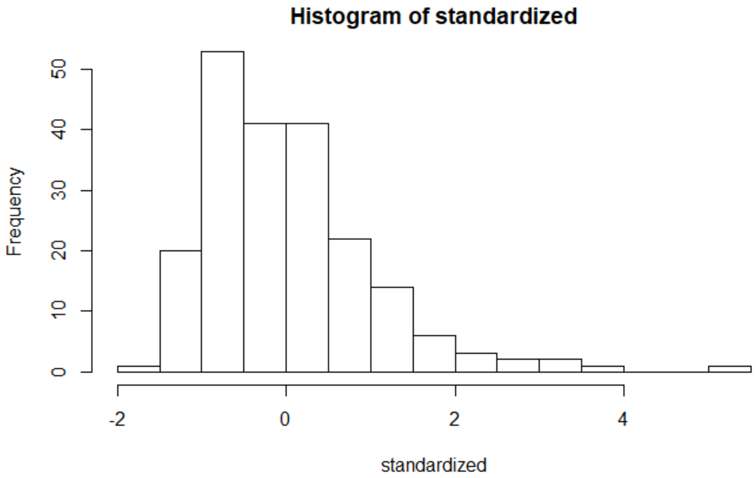
1. **Multivariate linearity**

   The assumption for linearity is met since the Normal Q-Q plot is nearly linear.



2. **Multivariate normality**

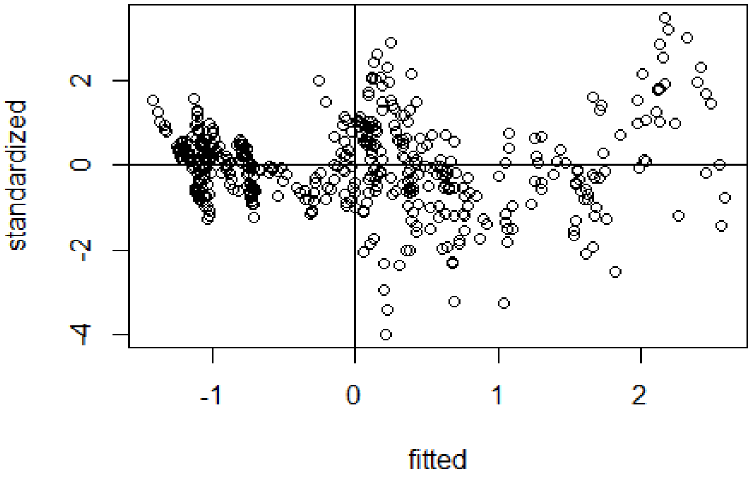
   The assumption for normality is met since the histogram presents a nearly normal distribution.



3. Homogeneity / Homoscedasticity

The assumption for homogeneity is met since the plot shows that the spread is nearly consistent across the ranges of values.

The assumption for homoscedasticity is met since the variance around the regression line is similar for all values of the predictor variable.



Conclusion

- Our best model is:

$IndexPr\widehat{ice}(monthly)$=*$430 - 6.8*Medicare - 1.12*Personal Interest Income + 0.53*Personal Dividend Income - 0.07*Real Disposable Personal Income + 32.8*Personal Savings Rate + 9.37*Personal Interest PMT + 1.04*Personal Current Taxes + 6.27*Social Security + ε*

- The movements of consumer-related indicators in our model are highly correlated with the monthly price change of S&P 500 index.
- Our dataset presents a limitation of our model. Most of our independent variables are correlated with each other, while they should not be multicollinear or colinear to meet the additivity assumption for the multiple regression.

References

1. FRED Economic Data: https://fred.stlouisfed.org/

2. Yahoo Finance: https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC