ISYE 6414, Spring 2022

Professor: Josh McDonald

ISYE 6414 Final Project Report

Authors:

Pranav Mehta

Rishi Dasgupta

Jinyu Park

Sara Alnasser

Date: 04/20/2022

Table of Content

1. Introduction

The primary purpose of this project is to accurately predict the returns of the S&P500 (Response Variable) by using a combination of Macro-variables (Predictor Variables). The goal is to explore the correlation between the returns of equity markets and the changes in macro-economic environment.

2. Motivation

In the financial world, there exists a strong consensus about the high correlation between the stock market and the overall health of the economy at large. In this regard, understanding the variability and correlation of market returns is an indispensable asset for formulating reliable investment strategies for such entities.
The purpose of this study is to utilize data mining and statistical analysis to predict the daily/monthly stock market (SP500) returns using Multiple Linear Regression. We will select ~20 macro-economic features that are significant to the economic health by literature review and use them to assess the predictive power of the variables and test the supposed correlation between the stock market and the economy. To make computation more compact, we will incorporate principal component analysis (PCA) to identify the best predictors in the model

3. Data Collection and Preprocessing

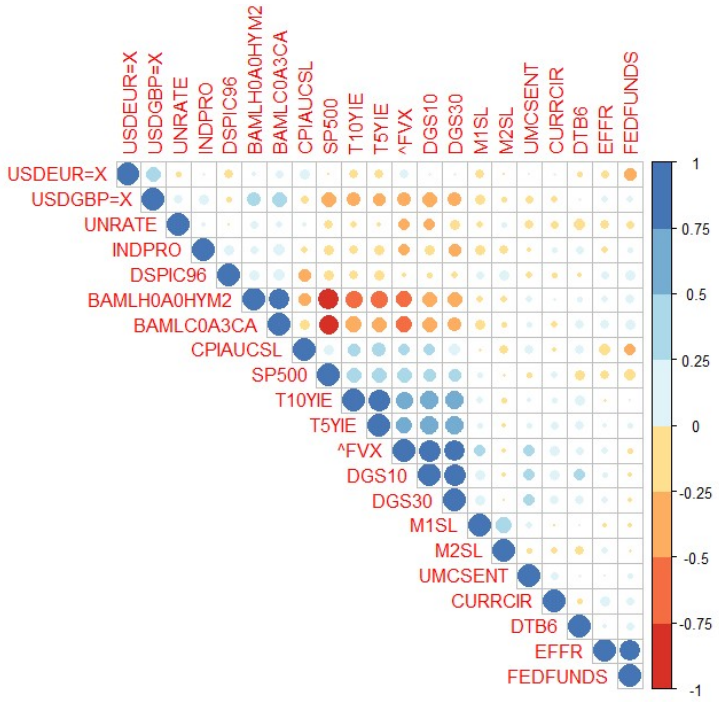
The accuracy of our project depends on the input quality of the data therefore we plan to collect our data using reliable websites. We will implement different in-built libraries in R such as “tidy verse” and ”fredr” to collect data from different websites.

The final data has 22 Macro variables in a Monthly Format. The final dataset is a monthly format ranging from January, 2015 to December, 2019. A few variables were collected in a daily format and few of them were collected in a monthly format. Thus we had to clean the data and match the frequency of the time. We converted the daily data by taking the mean of the daily changes to convert them to a monthly format. Some of the data collected from the websites had a lot of missing information, so we used a Backfill and Front fill to complete the dataset.

Once we had the monthly values, we calculated the monthly change in the variables as this is a better predictor for Regression Models. We also calculated the log returns to see if it was better transformation but the initial models were highly inaccurate so we shifted to monthly returns.

4. Exploratory Data Analysis

Before, fitting our model we wanted to understand the underlying predictive power of the independent variables used in the initial dataset. We plotted a correlation matrix to explore the correlated nature of the variables.



From the above plot we can see that the S&P500 returns are highly positively correlated with "T10YIE", "T5YIE", "^FVX" and "DGS10" and moderately negatively correlated with "EFFR" and "FEDFUNDS". After plotting the correlation matrix we expected that there might positive coefficient and negative coefficients respectively for the above variables and ~0 for the other variables. The correlation matrix did confirm that using the collected data, would be a good initial dataset to fit Multiple Linear Regression Models.

The above correlation plot also showed that a lot of Independent variables were highly correlated, thus we tried to reduce the variables in our models by using several Variable Selection Models to fit the best model.

5. Variable Selection

We initially fit the Full Model, to our initial dataset and calculated the VIF Index for the variables. We wanted to reduce the number of variables used thus we eliminated a few variables that had a high VIF Score.

```
> vif(model)
      DGS10        DTB6       DGS30 BAMLH0A0HYM2   BAMLC0A3CA       T10YIE        EFFR       T5YIE   `USDEUR=X`
  41.898934    1.678533   25.951092    7.168813    5.017658   11.206898    4.174001   11.652850    2.869968
`USDGBP=X`      `^FVX`      INDPRO     CPIAUCSL      UNRATE     FEDFUNDS     UMCSENT     DSPIC96        M1SL
   2.317006   11.445169    1.676360    1.705496    1.388502    4.154470    1.438776    1.537536    1.923848
    CURRCIR        M2SL
   1.428410    1.937371
```

After an initial VIF screening, the following 5 variables were removed due to the presence of high degree of multicollinearity: DGS10, DGS30, T10YIE, T5YIE, and `^FVX`.

We implemented variable selection on the 15 remaining predictor variables using the methods stated below and compared their model performance using adjusted R-squared and the F-statistic as the metrics for comparison.

- **Forward stepwise regression** with AIC criterion: Performed stepwise regression starting from a base model and adding variables that improve the model most, one at a time such that the Akaike Information Criterion (AIC) value is minimized.
- **Backward stepwise regression** with AIC criterion: Performed backward regression from a saturated (full) model by removing predictors based on Akaike Information Criterion (AIC), in a stepwise manner until a reduced model is produced which best explains the data.
- **Mallow's Cp:** Mallow's Cp compares the precision and bias of the full model with a subset of the predictors. We used this criterion to iteratively compare models with different subsets of predictor variables.
- **Ridge Regression**: Ridge regression shrinks the coefficients of the variables in the model (L2 regularization) and is especially good at improving least-squares estimate when multicollinearity is present. We performed ridge regression with bias coefficients varying from 0 to 10 in steps of 0.25.
- **Lasso Regression**: Lasso regression uses shrinkage to produce sparse models (reduces coefficients of less relevant variables to 0). It uses L1 regularization which adds penalty equal to the absolute value of the magnitude of the coefficients resulting in the elimination of some variables from the model.
- **Elastic Net Regression**: Elastic net regression is a regularized regression method that linearly combines the L1 and L2 penalties of the Lasso and Ridge regression methods. A bias coefficient of 0 converts elastic net to ridge regression and a value of 1 converts it to Lasso regression. We used an alpha value of 0.5 to get a midway model between lasso and ridge regression.

The above mentioned variable selection methods and model outputs have been explained in the next section.

6. Model Fitting, Results and Analysis

PCA Analysis

Principal Component Analysis (PCA) is a dimension reduction method that uses the process of computing principal components and utilizing them to perform a change of basis on the data. These principal components are essentially a newly constructed direction, or more simply a linear combination of the features. Each principal component is orthogonal to one another, thereby representing individual dimensions of the data that are linearly uncorrelated. There can be as many principal components as the original set of features at maximum, and the principal components are in order of variance explained

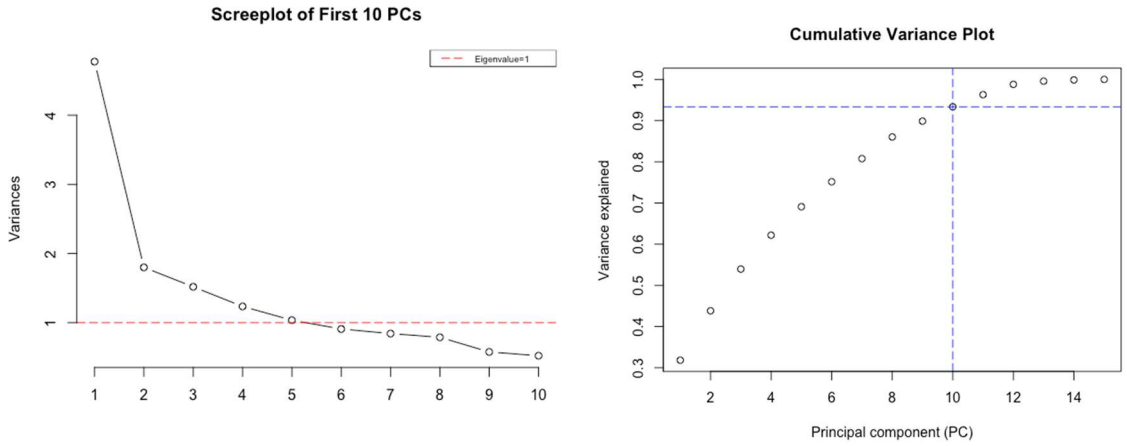


Figure 1 is a scree plot of the first ten principal components. A scree plot is a line plot of eigenvalues of principal components in the analysis. A factor, or a principal component, with an eigenvalue of 1 account for as much variance as a single variable and the logic is that only factors that can explain at least the same amount of variance as a single variable is worth keeping.

Furthermore, the cumulative variance by every increase of principal components was plotted to investigate how many principal components would be sufficient to represent the whole data dimension, preferably less than the 15 features originally used. Figure 2 shows the plot of cumulative variances at each one of 15 principal components. More than 90% of the variability in data is explained by the first ten principal components, suggesting that the data set could be effectively reduced from 15 dimensions to 10 dimensions using these ten principal components.

Forward, Backward, Mallows Cp and PCA Variable Selection

	model_variable_selections	r_squared	adjusted_r_squared
1	Multiple Linear	0.7659256	0.6458874
2	Forward Selection	0.7030555	0.6814595
3	Backward Selection	0.7236694	0.6923867
4	Mallows Cp	0.7030555	0.6814595
5	PCA	0.8525193	0.8224212

We then used several variable selection models and compared the Adjusted R squared of these models to find the best method to select the Predictor Variables. From our analysis we concluded that using a PCA analysis as Variable reduction tool gave us the best R-squared and Adjusted R-squared., followed by Backward Selection. We also observed that except PCA most other variable reduction methods had similar R-squared.

The most common variables in all model are as follows:

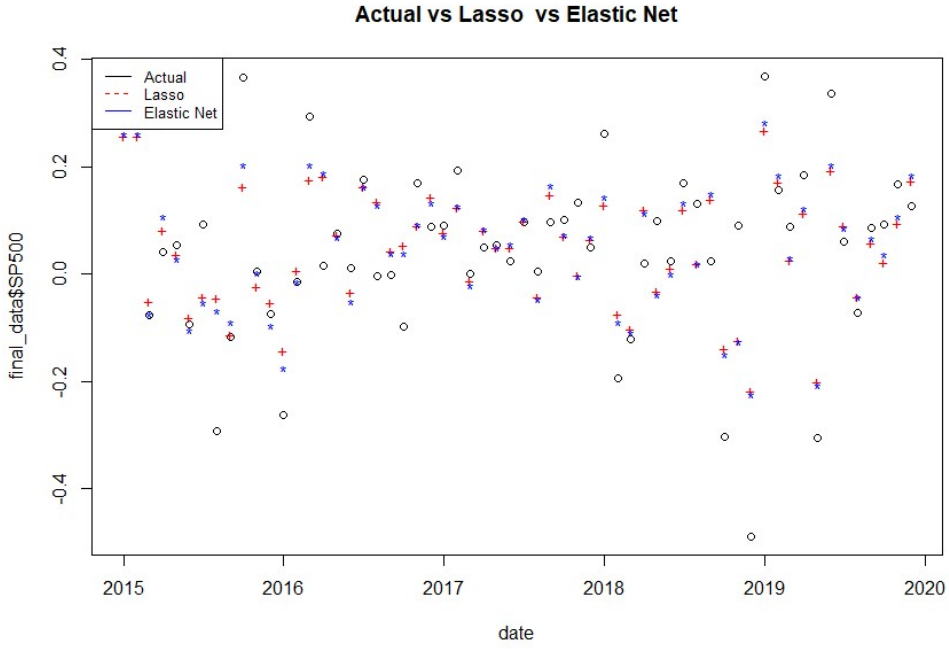
1. **DTB6**: 6-month Treasury Bill Secondary Market Rate
2. **BAMLH0A0HMYM2**: ICE Bank of America US High Yield Index value
3. **EFFR**: Effective Federal Funds Rate
4. **INDPRO**: Industrial Production Rate
5. **CURRCIR**: Currency in Circulation (USD)

6. **M2SL**: M2 index (Billions of Dollars)

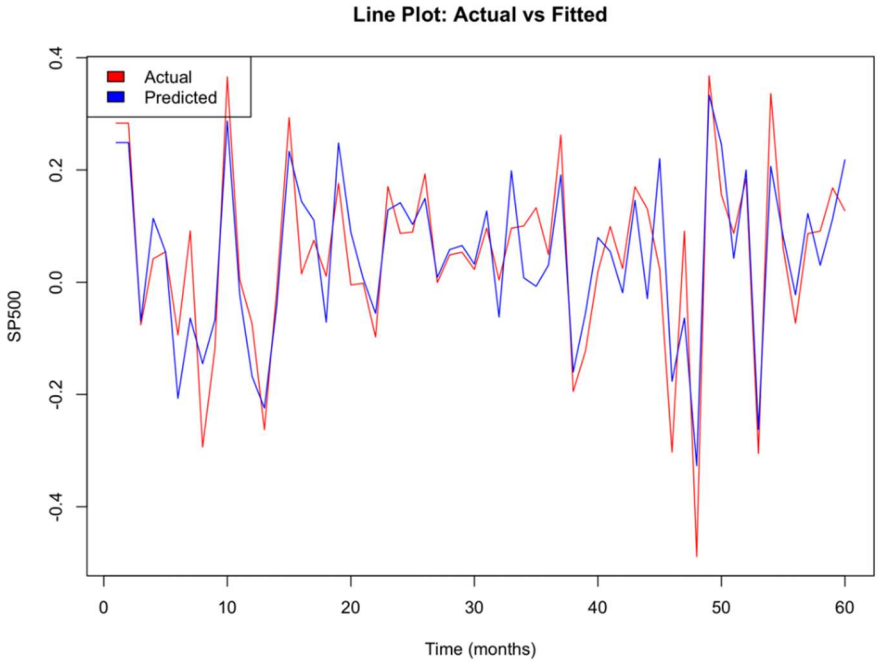
We can conclude that most of these variables have a different macro-economic importance and it is important to look at all aspects of the economy to predict the Equity Market returns.

Lasso, Ridge and Elastic Net Regression

We also used different regression models that penalize additional variables to reduce the variance-bias tradeoff.



	coef_names	lasso_coef	ridge_coef	elastic_coef
1	DTB6	0.00000000	-0.1287	-0.0059298897
2	BAMLH0A0HYM2	-0.18017583	-0.5554	-0.1743834254
3	BAMLC0A3CA	-0.08521304	-0.1838	-0.1028854426
4	EFFR	0.00000000	-0.1006	0.0000000000
5	USDEUR=X	0.00000000	0.0367	0.0000000000
6	USDGBP=X	0.00000000	-0.0809	-0.0449549412
7	INDPRO	0.00000000	-0.0780	0.0000000000
8	CPIAUCSL	0.00000000	-0.0270	0.0000000000
9	UNRATE	0.00000000	-0.0824	-0.0003643004
10	FEDFUNDS	0.00000000	-0.0189	-0.0003583574
11	UMCSENT	0.00000000	-0.0359	0.0000000000
12	DSPIC96	0.00000000	-0.0034	0.0000000000
13	M1SL	0.00000000	-0.0134	0.0000000000
14	CURRCIR	0.00000000	0.0790	0.0008318160
15	M2SL	0.00000000	-0.0944	-0.0105172414

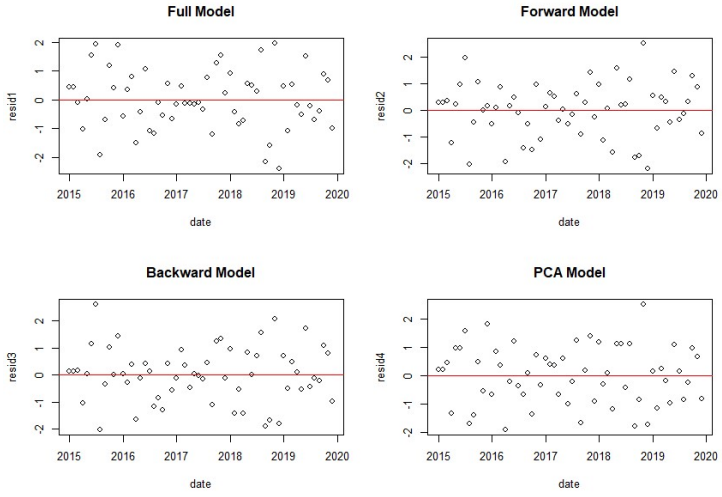


From the above plot, we can observe that Lasso and Ridge regression did a similar job at predicting the returns of S&P500, and there is conclusive evidence to chose one over the other. We can also conclude that these models gave higher importance to "BAMLH0A0HYM2" and "BAMLC0A3CA". These variables were not selected in the any of the variable reduction methods. We can conclude from the above figures that these regression models do a good job at predicting the Response variable and control the Bias-Variance Tradeoff.

7. Assumptions/ Model Validation

All the variable selection methods eventually produce multiple linear regression models, so the model assumptions to check for are common: the mean zero assumption, constant variance assumption, and normality assumption.

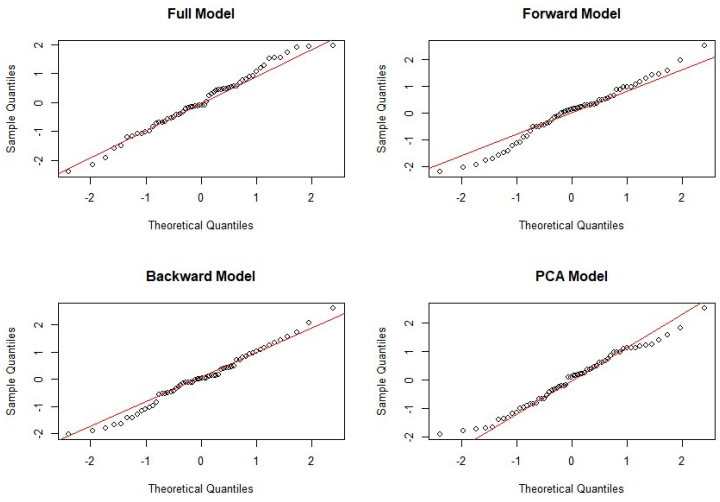
Linearity Assumption



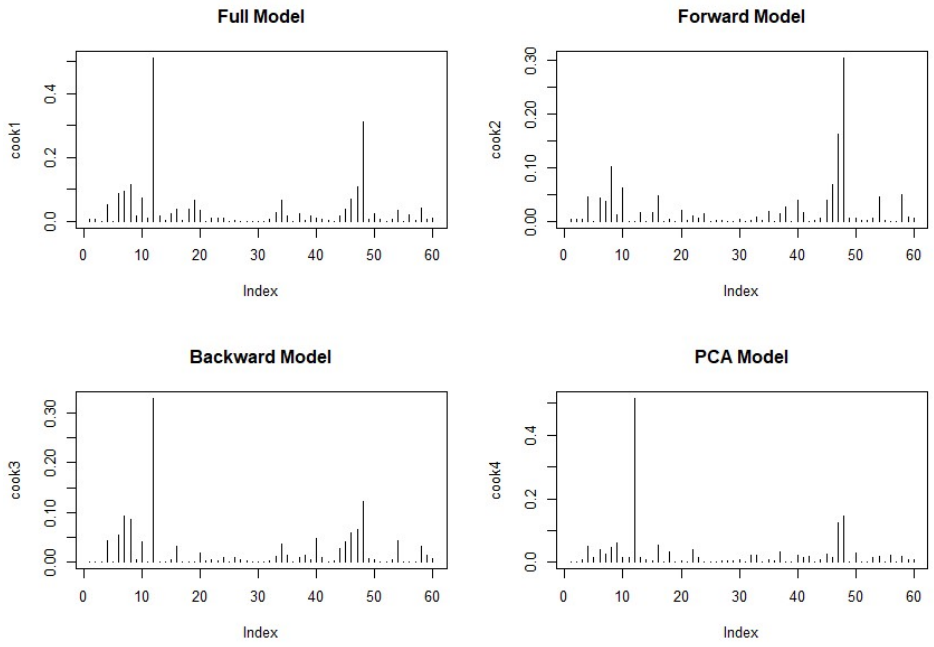
To check for mean zero assumption, we need to look at whether the standard residuals are randomly distributed within equal distance from zero. Figure 3 shows standard residual plots of the four regression models using different variable selection methods. Across all plots in Figure 3, we observed random distribution of standardized residual values within equal distances from mean zero. Moreover, from observing the absence of any clusters or patterns in the distribution of residual values, it is safe to conclude that constant variance assumption is fulfilled as well.

Normality Assumption

We also inspect quantile-quantile (Q-Q) plots to check for the normality assumption. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles are identically distributed, one should see the points roughly aligned along a straight line. Figure 4 shows a collection of Normal Q-Q plots of four regression models: the full multiple linear regression model using all 20 features, two models constructed from forward and backward stepwise regression methods, and finally the principal component regression model using 10 principal components that explain more than 90% of variability in the data.
The full MLR model and backward stepwise regression model produce Normal Q-Q plots that have points forming a line that is quite straight. However, forward stepwise regression model and principal component regression model produce Q-Q plots in which the tails seem to deviate significantly from the outline of a straight line. Although the tails in Normal Q-Q plots generally deviate from a straight line, this observation may suggest the presence of outliers in the data. Cook's Distance is computed in order to further investigate what is happening at the deviations.



Outliers



Cook's Distance is a metric to identify outliers in the data. In this case, we compute Cook's Distance using predicted values from the four regression models following this equation:

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})^T(\hat{Y}_{(i)} - \hat{Y})}{(k+1)\hat{\sigma}^2}$$

Fitted values from the model is compared with the model fitted values excluding the i^{th} observation, so a large Cook's Distance indicates a big change in estimated parameter values in the regression model by the removal of the i^{th} observation.

The above figure presents four plots of Cook's Distances from the four regression models of comparison. The rule of thumb for assessing whether a large Cook's Distance is truly an outlier is whether it is larger than 1. None of the large Cook's Distances across all regression models are greater than 1, therefore we conclude that there are no outliers and that the normality assumption is fulfilled.

8. Conclusion

The majority of the models fitted after variable selection performed adequately and approximately explained around 65% of the variability in the value of the S&P 500 index over the 5-year time period in consideration.
Among the best performing models were generated from the backward regression algorithm and the PCA method, with the latter explaining almost 83% of the variability in the model although the PCA model did not account for the bias-variance trade-off. In case we require the consideration of bias-variance trade-off, the lasso method for eliminating less relevant predictor variables is another good option.
In conclusion, we can confidently attest to the fact that the selected predictor variables in the final model best explain the variability in the S&P 500 index values over the time-period considered.

9. Further Research

1. Measuring prediction accuracy over different timeframes:

The multiple regression models in this project utilized records of predictor variables over a 5-year period from January 2015 to December 2019 which was identified as a period of relatively low volatility in the stock market.
To measure long-term trends, the timeframe may be increased over decades and the selected predictor variables may be compared with short-term predictors to quantify macro-economic trends that dictate long term market health.

2. Measuring prediction accuracy over specific periods of higher volatility:

Focusing on periods of higher volatility and identifying variables which impact short-term market rebounds such as those seen during the post-housing crisis and post-pandemic periods in 2009 and 2020 respectively could allow us to quantify the most impactful factors lawmakers need to consider during market recovery legislations.
Performing time-series analyses and ARIMA modelling over these volatile periods and comparing the actual values with the predicted values could allow us to quantify the short-term deviations presented due to unprecedented world events.

3. Utilizing different classification models:

Machine learning methods such as K-Nearest Neighbors and Hidden Markov Models can be used to classify market volatility into different regimes in-order to accurately predict switches between bullish and bearish characteristics.
Support Vector Machines can be used to forecast the movement of the S&P500 index. The inputs can be technical analysis indicators such as the Moving Average Convergence Divergence (MACD) and the Relative Strength Index (RSI). SVMC is used to optimize the values of the MACD and RSI to determine the best situations to buy or sell the index

10.Appendix

```
> summary(model)

Call:
lm(formula = SP500 ~ ., data = final_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.196997 -0.050790 -0.008652  0.047346  0.159809

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0575053  0.0498523   1.154   0.2557
DGS10         0.1591107  0.1794144   0.887   0.3806
DTB6         -0.0139918  0.0112737  -1.241   0.2220
DGS30        -0.1109156  0.1905785  -0.582   0.5639
BAMLH0A0HYM2 -0.3351907  0.0764529  -4.384 8.54e-05 ***
BAMLC0A3CA   -0.0245543  0.0739162  -0.332   0.7415
T10YIE        0.0991028  0.1258158   0.788   0.4356
EFFR         -0.0060954  0.0195310  -0.312   0.7566
T5YIE        -0.1270045  0.1046144  -1.214   0.2320
`USDEUR=X`    0.2538781  0.2514692   1.010   0.3189
`USDGBP=X`   -0.1563522  0.1606024  -0.974   0.3363
`^FVX`       -0.1251997  0.0791819  -1.581   0.1219
INDPRO       -0.0601603  0.0329229  -1.827   0.0753 .
CPIAUCSL     -0.0513291  0.1085493  -0.473   0.6389
UNRATE       -0.0027647  0.0056551  -0.489   0.6277
FEDFUNDS     -0.0001316  0.0016963  -0.078   0.9385
UMCSENT       0.0016476  0.0046834   0.352   0.7269
DSPIC96       0.0244150  0.0659969   0.370   0.7134
M1SL          0.0111365  0.0222051   0.502   0.6188
CURRCIR       0.0858464  0.0519744   1.652   0.1066
M2SL         -0.1296895  0.0803798  -1.613   0.1147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09889 on 39 degrees of freedom
Multiple R-squared:  0.7659,    Adjusted R-squared:  0.6459
F-statistic: 6.381 on 20 and 39 DF,  p-value: 4.385e-07
```

```
> summary(backward_model)

Call:
lm(formula = SP500 ~ DTB6 + BAMLH0A0HYM2 + EFFR + INDPRO + CURRCI
    M2SL)

Residuals:
      Min        1Q    Median        3Q       Max
-0.175490 -0.045055  0.001423  0.061048  0.229354

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.048926   0.036387   1.345   0.1845
DTB6         -0.015471   0.008329  -1.857   0.0688 .
BAMLH0A0HYM2 -0.288336   0.027109 -10.636 9.31e-15 ***
EFFR         -0.017275   0.009339  -1.850   0.0699 .
INDPRO       -0.039166   0.024573  -1.594   0.1169
CURRCIR       0.074764   0.042253   1.769   0.0826 .
M2SL         -0.081609   0.055676  -1.466   0.1486
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09217 on 53 degrees of freedom
Multiple R-squared:  0.7237,    Adjusted R-squared:  0.6924
F-statistic: 23.13 on 6 and 53 DF,  p-value: 3.328e-13
```

```
> summary(forward_model)

Call:
lm(formula = SP500 ~ DTB6 + BAMLH0A0HYM2 + BAMLC0A3CA + M2SL,
    data = final_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17829 -0.04757  0.01431  0.05102  0.22245

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.075468   0.028619   2.637   0.0109 *
DTB6         -0.016309   0.008514  -1.915   0.0606 .
BAMLH0A0HYM2 -0.227471   0.049362  -4.608 2.45e-05 ***
BAMLC0A3CA   -0.098000   0.057477  -1.705   0.0938 .
M2SL         -0.085465   0.055636  -1.536   0.1302
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09379 on 55 degrees of freedom
Multiple R-squared:  0.7031,    Adjusted R-squared:  0.6815
F-statistic: 32.55 on 4 and 55 DF,  p-value: 6.409e-14
```

```
> summary(pca_model)

Call:
lm(formula = SP500 ~ ., data = principal_components)

Residuals:
      Min        1Q    Median        3Q       Max
-0.123706 -0.051766  0.007736  0.047747  0.159759

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.046678   0.009041   5.163 4.41e-06 ***
PC1         -0.046162   0.003767 -12.255  < 2e-16 ***
PC2         -0.041274   0.005787  -7.132 4.15e-09 ***
PC3          0.028437   0.006439   4.416 5.53e-05 ***
PC4         -0.028931   0.007200  -4.018 0.000202 ***
PC5          0.040501   0.007777   5.208 3.78e-06 ***
PC6         -0.020534   0.008268  -2.484 0.016479 *
PC7          0.018675   0.008695   2.148 0.036706 *
PC8         -0.020095   0.009154  -2.195 0.032920 *
PC9          0.000681   0.009878   0.069 0.945318
PC10        -0.020793   0.010623  -1.957 0.056013 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07003 on 49 degrees of freedom
Multiple R-squared:  0.8525,    Adjusted R-squared:  0.8224
F-statistic: 28.32 on 10 and 49 DF,  p-value: < 2.2e-16
```

```
> summary(mallows_cp_model)

Call:
lm(formula = SP500 ~ DTB6 + BAMLH0A0HYM2 + BAMLC0A3CA + M2SL,
    data = final_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17829 -0.04757  0.01431  0.05102  0.22245

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.075468   0.028619   2.637   0.0109 *
DTB6         -0.016309   0.008514  -1.915   0.0606 .
BAMLH0A0HYM2 -0.227471   0.049362  -4.608 2.45e-05 ***
BAMLC0A3CA   -0.098000   0.057477  -1.705   0.0938 .
M2SL         -0.085465   0.055636  -1.536   0.1302
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09379 on 55 degrees of freedom
Multiple R-squared:  0.7031,    Adjusted R-squared:  0.6815
F-statistic: 32.55 on 4 and 55 DF,  p-value: 6.409e-14
```