STA 141A Final Project

Ruhaan Juyal, Sia Puri, Tiffany Tran, Zhehao Zhang

2024-06-12

Introduction

The S&P 500 index, maintained by S&P Global, represents the performance of the 500 largest publicly listed companies in the United States. As a barometer of the overall health of the US stock market and a popular investment tool, predicting the annual returns of the S&P 500 can provide valuable insights for investors and policymakers. This project aims to explore the relationship between key macroeconomic indicators and the historical performance of the S&P 500, ultimately developing a predictive model to guide future investment strategies.

The economic indicators this report will utilize are:

Inflation rate: This is a measure of the average percentage increase in prices over the year. It is often regarded as a measure of an economies health, as high inflation reduces the capacity of consumers to spend on goods and services which harms the overall economy.

Unemployment rates: This is a measure of the percentage of the working age population that is looking for but cannot find a job. Increases in this measure are linked with reduced output and spending, which in turn worsen the performance of various industries.

Federal funds rates: This is the interest rate charged by the Federal Reserve system for banks to borrow money. It is also the target rate set by the Reserve's board of governors for the economy at large. This means that it is a accurate indicator of the price that businesses and consumers pay to borrow money for investment.

GDP growth rates: This is the annual percentage growth of the Gross Domestic Product (GDP) in the United States. The GDP is a measure of the total dollar value of all goods and services produced in the US over the calendar year, meaning its growth rate is a direct measure of economic growth.

Recession: This variable will be categorical, stating whether a year was a recession (meaning the GDP declined for two straight quarters) or non-recessionary. This variable is closely linked to GDP Growth but targets a specific aspect of the measure (namely highlights any sustained declines).

Understanding how these factors influence the S&P 500 returns can shed light on broader economic dynamics and inform decision-making processes. We will investigate the extent to which these indicators correlate with S&P 500 performance and identify the combination of predictors that yields the most accurate model.

This report will utilize a dataset sourced from both the Federal Reserve and the World Bank which encompasses annual data for the named variables from 1961 onwards. We will apply a linear regression model to determine the predictive power of the selected macroeconomic variables. The dataset includes the annual S&P 500 return, annual inflation rate, annual average unemployment rate, annual average federal funds rate, annual GDP growth rate and a categorial recession variable.

By examining and evaluating several models constructed with combinations of these variables we aim to address the following questions: **Are macroeconomic indicators correlated with the performance of**

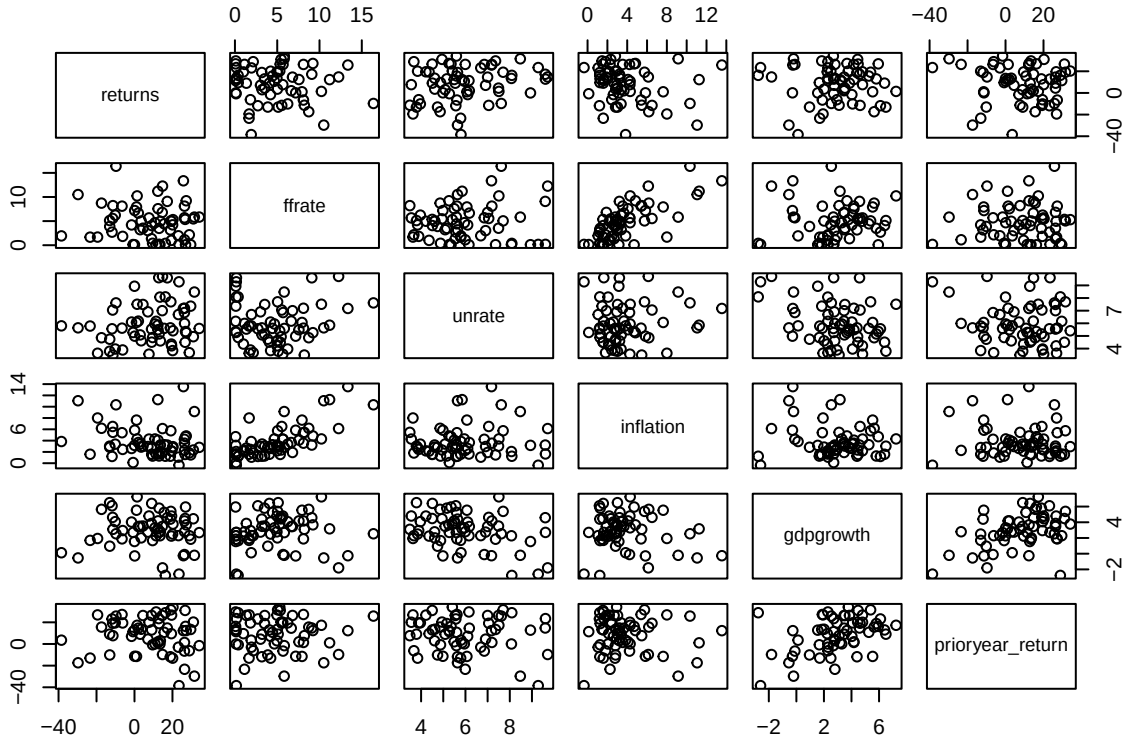
the S&P 500? What combination of indicators produces the most accurate model? Does the most accurate model produce predictions that are reliable enough for practical use?

We hope that by addressing these questions, our project will contribute to the understanding of the interplay between economic factors and stock market performance, offering insights that can enhance investment strategies and economic policy formulation.

The methodology for addressing this question will involve several stages. Initially, we will perform exploratory data analysis and analyze the correlation between each predictor and the S&P 500 return to inform our model variable selection. Subsequently, we will fit regression model using backwards selection, compare our results with cross-validation methods, and create visualizations using ggplot to illustrate the relationships between predictors and the response variable. Finally, we will conduct model diagnostics to ensure the robustness of our regression model, examining constant variance, independent error normality, and the presence of outliers.

Exploratory Data Analysis

Before diving into model fitting, we should do exploratory data analysis first to see if there are any relationships between variables we can examine visually. This can help identify likely and unlikely predictors for our response variable, while also showing what relationships predictors may have with one another. For our quantitatively-measured variables, we can easily see all these possible relationships with a pairs plot, as seen below.



From looking at these scatter plots, there does not seem to be clear, strong visual relationships between S&P 500 returns and most of these variables. And due to this uncertainty, it is also unclear whether these relationships are necessarily linearly related. However, this is not surprising as market returns are known to be unpredictable, so weaker relationships between possible predictors can actually be seen as reasonable for our model. And for the purposes of our model building, we will view these relationships as possibly linear, as there is no visual indication to show otherwise.

Federal funds rate and prior year returns could possibly have negative linear relationships with S&P 500

returns, while unemployment rate and GDP growth could have weak positive linear relationships instead. But as mentioned before, these relationships are very vague, and their significance in the model should be more closely and formally examined through other methods. On the other hand, the final variable, inflation, does seem to have a moderately strong, negative linear relationship with S&P 500 returns. Thus, its inclusion in the model is highly likely, especially in comparison with the other possible predictors.

Looking at the graphs between the predictors themselves, some relationships stand out as possibly significant. Federal funds rate and inflation appear to have a strong positive linear relationship with one another, so these two variables are likely collinear. Similarly, there may also be a negative linear relationship between prior year returns and inflation, although this appears to only be moderately strong. Notably, GDP growth seems to have multiple visible relationships with federal funds rate, inflation, and prior year returns. Thus this suggests GDP growth may have multicollinearity with these variables and may not be a relevant predictor in our model due to the inclusion of these other predictors. These possible collinear relationships should be kept in mind during the model fitting process when considering which predictors to keep or not.

Value (1 = Recession Year)	Count
0	48
1	14

Since we are also possibly including a categorical predictor in our model, it is also important to observe what the distribution of this variable is. Here, the categories of "recession year" and "not a recession year" are represented using the dummy variables of "1" and "0" respectively. This is done to translate these categories into numerical values so that we can appropriately fit them in our model. Since this category is binary, we can easily see its distribution through a simple table of counts as shown above. Here, it is clear that only a small proportion of years, only about a fifth of them, was considered to have a recession period.

```
##                 returns ffrate unrate inflation gdpgrowth prioryear_return recession
## returns            1.00  -0.09   0.22     -0.21     -0.01            -0.09     -0.17
## ffrate            -0.09   1.00   0.08      0.72      0.07             0.01      0.31
## unrate             0.22   0.08   1.00      0.08     -0.34            -0.12      0.29
## inflation         -0.21   0.72   0.08      1.00     -0.14            -0.07      0.38
## gdpgrowth         -0.01   0.07  -0.34     -0.14      1.00             0.44     -0.66
## prioryear_return  -0.09   0.01  -0.12     -0.07      0.44             1.00     -0.32
## recession         -0.17   0.31   0.29      0.38     -0.66            -0.32      1.00
##
## n= 62
##
##
## P
##                 returns ffrate unrate inflation gdpgrowth prioryear_return recession
## returns                 0.4943 0.0792 0.1077    0.9169    0.4817           0.1948
## ffrate          0.4943         0.5409 0.0000    0.5992    0.9142           0.0150
## unrate          0.0792  0.5409        0.5165    0.0070    0.3608           0.0217
## inflation       0.1077  0.0000 0.5165           0.2644    0.5744           0.0023
## gdpgrowth       0.9169  0.5992 0.0070 0.2644              0.0003           0.0000
## prioryear_return 0.4817 0.9142 0.3608 0.5744    0.0003                     0.0104
## recession       0.1948  0.0150 0.0217 0.0023    0.0000    0.0104
```

Beyond looking at graphs, the calculated correlation matrix can also provide a lot of information on the best potential models we can fit. The unemployment rate, inflation rate, and recession all appear to have reasonably high correlations with the return of the S&P 500 with low p-values. While they are not individually significant at a 0.05 threshold, their low correlations to each other would suggest they do not run into collinearity issues. As a result, a model with these varibles should produce a significant result.

On the other hand, GDP growth has an extremely low correlation and a very high p-value to match, so we will exclude it from further testing as it does not provide useful information in predicting the response variable. It also has high correlations with other predictors, which matches with our earlier observations with the graphs from before. Both federal funds rate and prior year returns neither have high or low p-values and

correlations with market returns, so they cannot be definitively excluded as predictors. However, it can be noted that federal funds rate has a very high correlation with inflation, which could cause collinearity issues in the model.

Linear Regression

After determining which variables are appropriate, we can fit the linear regression model. First, we can fit the full model, one which includes all of our variables except the excluded GDP growth, and see which variables are significant. After, we can use backwards selection to compare models with and without the least significant predictors until we get our final model. This final model should have the highest adjusted R-squared, which is a measure of model fit that explains how much of the variation of the dependent variable can be explained by the independent variables. Additionally, unlike R-squared, which is a very similar statistic, it adjusts itself for different numbers of predictors so that overfitting can be avoided.

```
##
## Call:
## lm(formula = returns ~ inflation + ffrate + unrate + prioryear_return +
##     recession, data = financial_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.088 -10.437   0.133  10.988  30.924
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -3.0547     8.3704  -0.365   0.7165
## inflation         -1.5237     1.0730  -1.420   0.1611
## ffrate             0.7019     0.8019   0.875   0.3852
## unrate             2.9215     1.2963   2.254   0.0281 *
## prioryear_return  -0.1628     0.1332  -1.222   0.2268
## recession         -9.7870     5.6350  -1.737   0.0879 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.65 on 56 degrees of freedom
## Multiple R-squared:  0.1612, Adjusted R-squared:  0.08626
## F-statistic: 2.152 on 5 and 56 DF,  p-value: 0.07251
```

As the p-values for unemployment rate and recession are less than our alpha level of 0.10, we can deduce that these are significant predictors in the model. However, the other predictors are not statistically significant. But as the model has an extremely small R-squared, overall these predictors account for a very low amount of variability in the response variable.

So instead, we can determine which variables may produce the best possible model, even if the results are not necessarily significant at typical statistical testing levels. As federal return rate seems to be largely less significant than the other predictors due to its high p-value, it may be possible to exclude it in the model. Additionally, we saw that it was highly correlated with inflation, which does have a smaller p-value.

Following our process of backwards selection, we can federal return rate as a predictor and fit another model with the remaining variables. Then, we can compare the results of this fitting with our previous model, which is shown in the table below. In this model, the adjusted R-squared value is greater than what it was in the last model, which suggests that this model produces a better fit against the response than the previous one. Now, prior year returns appears to be the most insignificant predictor out of those remaining in the model. It did have a similar correlation with S&P 500 returns to federal return rate, which suggests that it may also be reasonable to exclude in the final model.

Predictors Used in Model	R-Squared	Adjusted R-Squared
inflation + ffrate + unrate + prioryear_return + recession	0.1611554	0.0862586
inflation + unrate + prioryear_return + recession	0.1496806	0.0900091
inflation + unrate + recession	0.1307189	0.0857561
unrate + recession	0.1094077	0.0792181

unrate	0.0504506	0.0346247

But as we can see from the table above, this new model, which now excludes prior year returns, has a smaller adjusted R-squared value than the model before. Therefore, it is probably best to stick with our previous model as it provides us with a better fit. Additionally, this shows why including some non-significant predictors at the 0.10 alpha level is likely for our particular model. And as seen in our table, if we continue to remove the least significant predictors we will get lower adjusted R-squared values. So based on this criterion, we can stop our model selection at 4 predictors.

Cross-Validation

However, as we built the model only based on the limited historical data we had, we cannot ensure that our model can make accurate predictions. This is especially important considering our dataset only has 62 observations, meaning it can easily be manipulated by outlier points or overfitted to specific trends in time.

So to verify that our model makes the best predictions, we can use cross-validation to find which model will have the lowest MSE. These methods split our data into different training and testing sets repeatedly, and calculate the average MSE of the model after training and testing on all possible combinations of the sets. Here, we will be using both k-fold cross-validation and leave-one-out cross-validation (LOOCV). But since our dataset is small, we have to consider that the k-fold approach may be greatly affected by randomness. However, we can test this by using different values for k, such as k=5 and k=10. And since LOOCV is a special case of k-fold, we can also see how models compare with no fold selection randomness.

MSE is another measure of fit, similar to R-squared, which captures the average of all the errors against our regression line. These two statistics measure the same idea, but have different representations and interpretations. For the case of cross-validation, we compare the average of the MSEs gotten by our different models. And since we want to minimize our errors, we choose the model with the smallest MSE.

Predictors Used in Model	K=5-fold	K=10-fold	LOOCV
inflation + ffrate + unrate + prioryear_return + recession	251.6875	274.0086	281.4741
inflation + unrate + prioryear_return + recession	253.7251	267.0173	273.5090
inflation + unrate + recession	252.8561	259.8179	269.9017
unrate + recession	248.5554	257.6766	261.4793
unrate	259.0412	261.9333	264.1930

The results of using our cross-validation methods can be seen in the table above. Clearly, the best model when using all three methods is the one that includes only 2 predictors: unemployment rate and recession years. Even with the inherent randomness of the k-fold methods, this model still had the smallest MSE. Notably, this model is different from the model we found using backwards selection. But since cross-validation is more reliable as it depends on repeated testing, we can conclude that this should be our final model.

However to further assess our findings, we can also test different information criteria. These other criteria are Our criteria are Log-Likelihood (LL), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Predicted Residual Sum of Squares (PRESS), and Adjusted R^2. The log-likelihood measures how well the model explains the observed data. Higher log-likelihood values indicate a better fit. It is used as a basis for calculating other criteria like AIC and BIC. The AIC is used to compare models by evaluating their goodness of fit while penalizing for complexity. Lower AIC values indicate a model that balances fit and simplicity effectively. Similar to AIC, BIC also balances model fit with complexity but imposes a stronger penalty for the number of parameters. Lower BIC values suggest a better model, with a stronger emphasis on model simplicity than AIC. PRESS assesses the model's predictive performance by measuring how well it predicts new data points. Lower PRESS values indicate a model that generalizes better to unseen data. R^2 Adjusted adjusts the R-squared value for the number of predictors in the model, providing a measure of the model's explanatory power that accounts for the number of predictors.

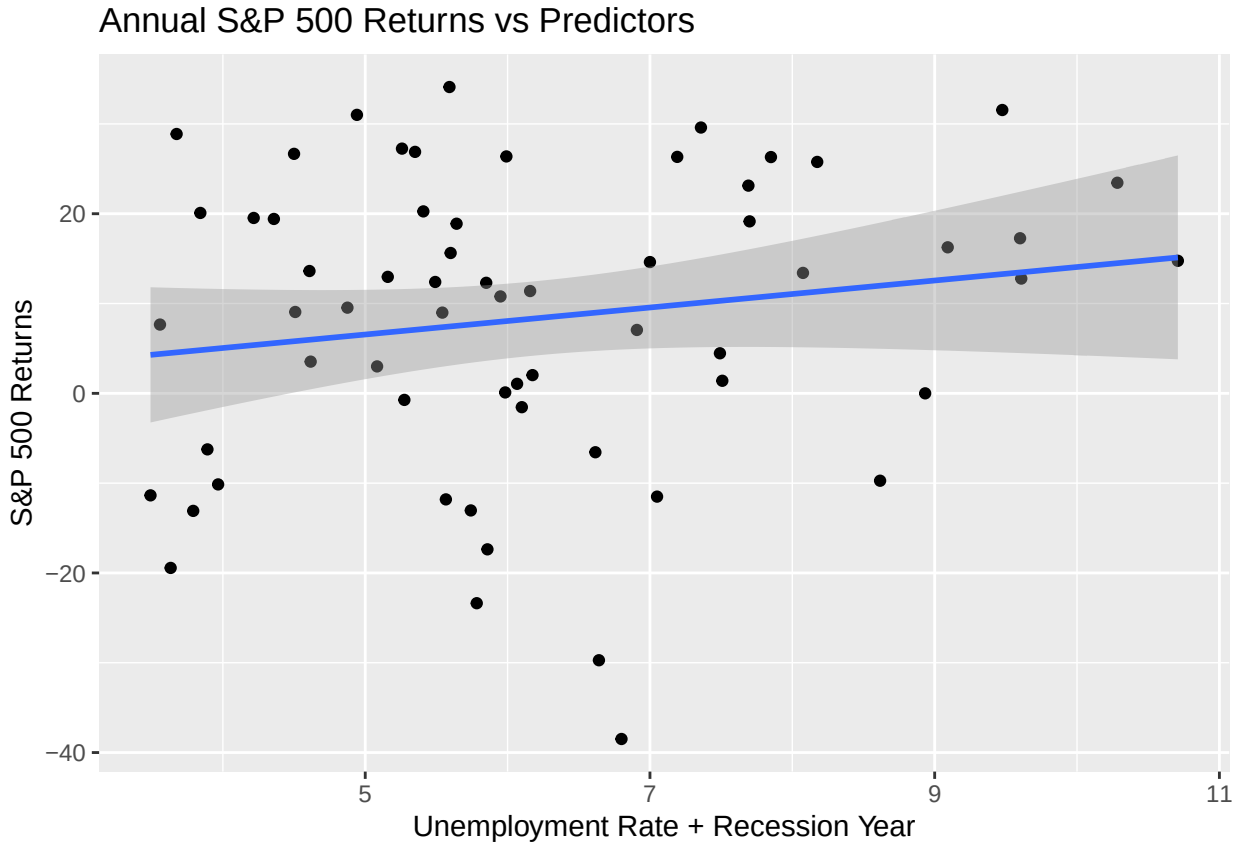
```
##                   p        LL      AIC      BIC    PRESS      R2adj
## y~x1+x2+x3+x4+x5 6 -255.3318 522.6637 535.4265 17451.39 0.08625861
## y~x1+x3+x4+x5    5 -255.7530 521.5060 532.1417 16957.56 0.09000910
## y~x1+x3+x5       4 -256.4367 520.8734 529.3819 16733.90 0.08575610
## y~x1+x5          3 -257.1875 520.3751 526.7565 16211.71 0.07921811
## y~x1             2 -259.1747 522.3493 526.6036 16379.97 0.03462473
```

The data in the table that show, model 4 has the smallest value of AIC and PRESS has the second smallest value of BIC, which suggest that Model 4 is the best model that we can choose. Thus, both of these two method indicate that model 4 is the best model to choose, which is as same as the LOOCV method and prove our model selection is correct

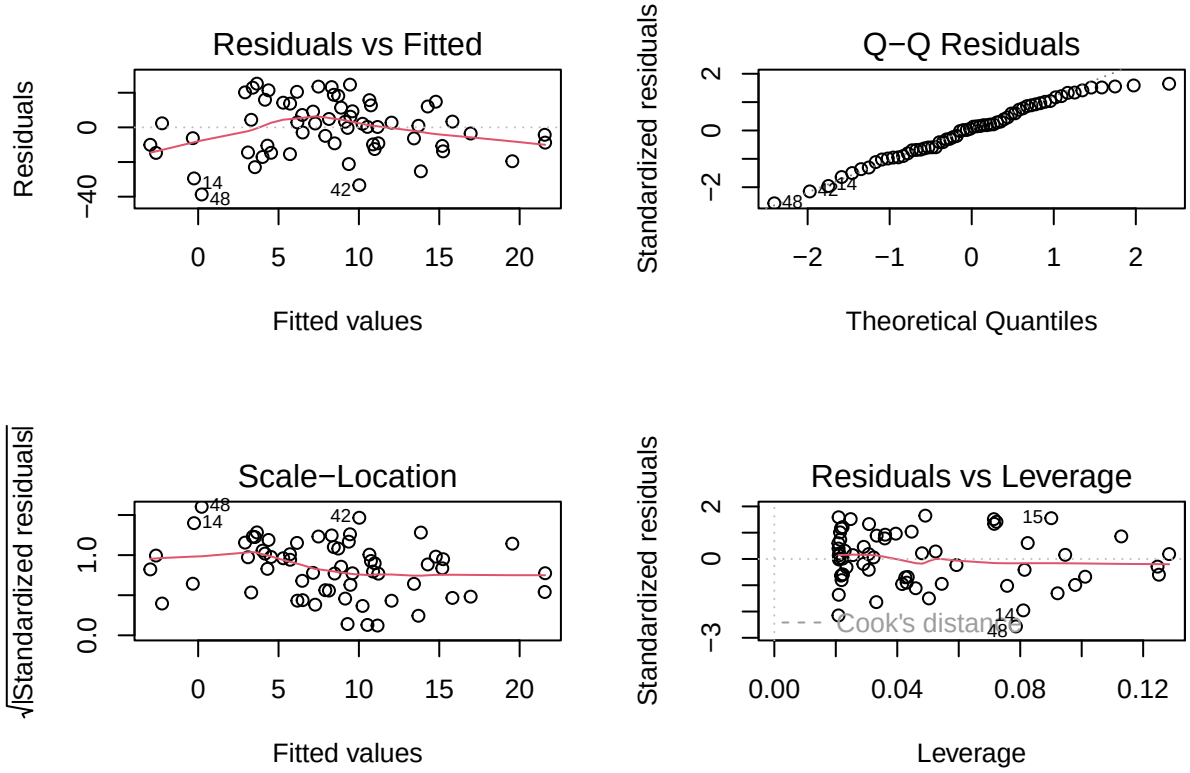
Model Evaluation

The final model that we choose is model 4 which is built by two predicted variable recession and unemployment rate, the response variable is the return.

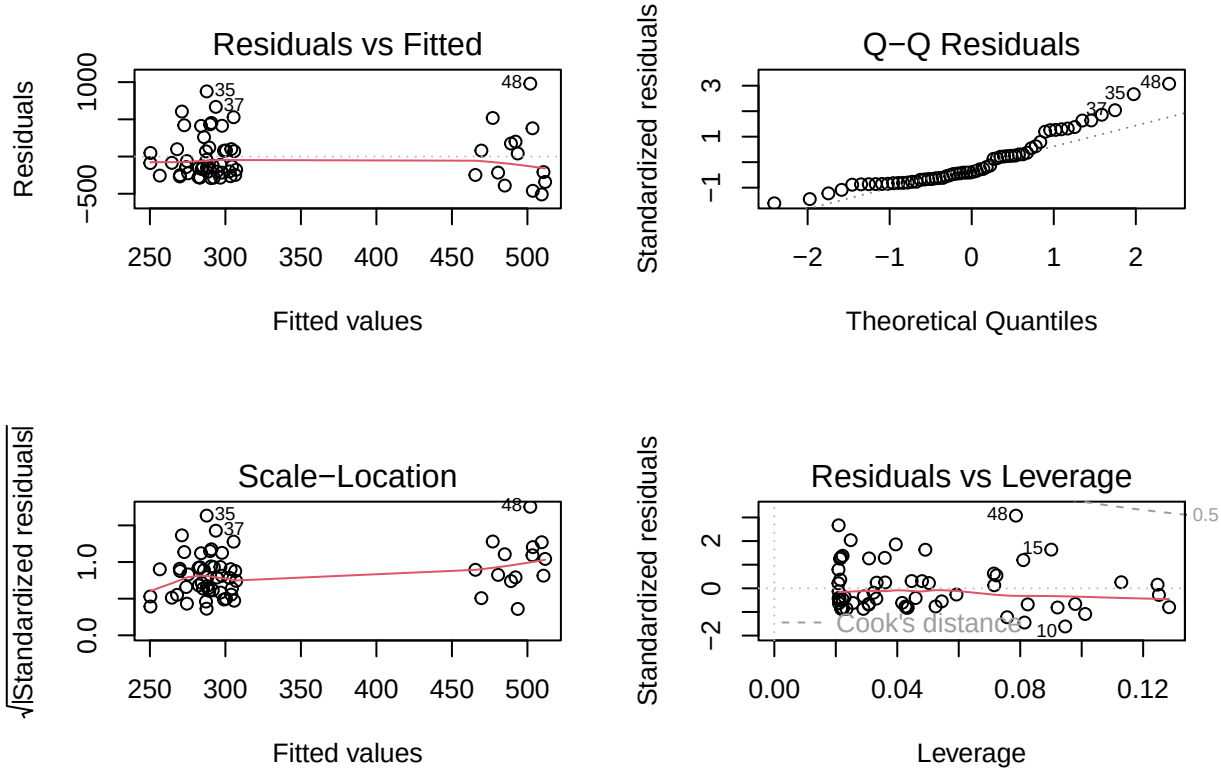
Next, we should create a plot to examine this relationship visually. Like the results of the model fitting process imply, these variables are not very strong predictors of market returns. However, there does appear to generally be a weak, positive linear relationship between the response and the predictors, as indicated by the blue line. It seems like the data points also vaguely follow this trend, rather than show nonlinearity.



Next step is to do linear regression model diagnostic to test if this model satisfy 4 assumption of linear regression model which are 1 error independence; 2 error in normal distribution;3 Constant variance in residual;4 show linear relationship.



As the result, we can see final model satisfy three out of four assumption,but in fitted value VS residual plot. The variance line show a upward curve which indicate there is no constant variance. Rest of the plot show all error are independently normally distributed, and show there is a linear relationship between predict variables and response variables. There are three high leverage points which are 14th,15th,and 48th observation, since our data sample's size is too small-only 62 so we decided to keep them, in case deleting might influence our model result. The method that fixing non constant variance we use Transformation of our response variable-return.



For the transformation part, we use square of the return. After the fixing, the model satisfy the assumption of the constant variance, even there are some point show non normal distribution in the QQ plot, but we think it is still in the acceptable range.

Conclusion

In this project, we aimed to predict the annual returns of the S&P 500 index by leveraging key macroeconomic indicators such as the inflation rate, unemployment rate, federal funds rate, GDP growth rate, and the presence of a recession. Our primary objective was to identify the variables that significantly influence S&P 500 returns and to develop a predictive model to aid investors and policymakers. Our initial exploratory data analysis suggested weak linear relationships between S&P 500 returns and the selected economic indicators. However, the unemployment rate, inflation rate, and recession status emerged as potential predictors due to their moderate correlations with S&P 500 returns. We also observed multicollinearity between certain predictors, particularly between the federal funds rate and inflation. The correlation analysis confirmed moderate correlations for the unemployment rate, inflation rate, and recession status with S&P 500 returns. Conversely, GDP growth showed a very low correlation with S&P 500 returns and high collinearity with other predictors, leading to its exclusion from further analysis. We then developed several linear regression models. The full model, which included all predictors except GDP growth, had low explanatory power, with an adjusted R-squared of 0.08626. Using backward selection, we excluded the federal funds rate due to its high p-value and collinearity with inflation. The best model, based on adjusted R-squared, included inflation, unemployment rate, prior year returns, and recession status. Cross-validation methods, including k-fold and LOOCV, indicated that the model with only the unemployment rate and recession status had the lowest Mean Squared Error (MSE), suggesting it as the most reliable model. Diagnostics on the final model indicated that three out of the four assumptions of linear regression were satisfied. The assumption of constant variance was initially violated but was corrected by transforming the response variable (return). The transformed model met all assumptions satisfactorily, demonstrating robustness in predicting S&P 500

returns. Our analysis indicates that the unemployment rate and the presence of a recession are the most significant predictors of S&P 500 returns among the variables studied. Although these predictors only explain a modest portion of the variance in S&P 500 returns, their inclusion in predictive models can still offer valuable insights for investment strategies and economic policy formulation. The findings suggest that while market returns are inherently unpredictable and influenced by a variety of factors, macroeconomic indicators such as unemployment rate and recession status can provide meaningful signals. In conclusion, our study contributes to the understanding of the interplay between economic factors and stock market performance, potentially enhancing investors' strategies and formulating economic policies.

Key Findings

`1. Exploratory Data Analysis:`

Initial visual analysis suggested weak linear relationships between S&P 500 returns and the selected economic indicators. The unemployment rate, inflation rate, and recession status showed potential as predictors due to their moderate correlations with S&P 500 returns. Multicollinearity was observed between certain predictors, particularly between the federal funds rate and inflation.

`2. Correlation Analysis:`

The correlation matrix confirmed moderate correlations for the unemployment rate, inflation rate, and recession status with S&P 500 returns. GDP growth showed a very low correlation with S&P 500 returns and high collinearity with other predictors, leading to its exclusion from further analysis.

`3. Linear Regression Models:`

The full model including all predictors (excluding GDP growth) had low explanatory power, with an adjusted R-squared of 0.08626. Using backward selection, the federal funds rate was excluded due to its high p-value and collinearity with inflation. The best model based on adjusted R-squared included inflation, unemployment rate, prior year returns, and recession status.

`4. Cross-Validation:`

Cross-validation methods (k-fold and LOOCV) indicated that the model with only the unemployment rate and recession status had the lowest Mean Squared Error (MSE), suggesting it as the most reliable model.

`5. Model Evaluation:`

Diagnostics on the final model indicated that three out of the four assumptions of linear regression were satisfied. The assumption of constant variance was initially violated but was corrected by transforming the response variable (return). The transformed model met all assumptions satisfactorily, showing it as a robust model for predicting S&P 500 returns.

Our analysis indicates that the unemployment rate and the presence of a recession are the most significant predictors of S&P 500 returns among the variables studied. Although these predictors only explain a modest portion of the variance in S&P 500 returns, their inclusion in predictive models can still offer valuable insights for investment strategies and economic policy formulation. The findings suggest that while market returns are inherently unpredictable and influenced by many factors, macroeconomic indicators such as unemployment rate and recession status can provide meaningful signals. Our study contributes to the understanding of the interplay between economic factors and stock market performance, potentially enhancing investors' strategies and formulating economic policies.

Code Appendix

```r
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
install.packages("Hmisc")
# Import data
returns = read.csv("returns.csv")
inflation = read.csv("inflation.csv")
unrate_monthly = read.csv("unrate.csv")
ffrate_monthly = read.csv("ffrate.csv")
gdpgrowth = read.csv("growth.csv")
recession = read.csv("recession.csv")
# Filter rows to start in 1961 and end in 2022
  # (except prior year returns which run from 1960 to 2021)
inflation = inflation[-1,]

returns_prioryear = returns[-(1:32),]
returns_prioryear = head(returns_prioryear, -3)

returns = returns[-(1:33),]
returns = head(returns, -2)

unrate_monthly = unrate_monthly[-(1:156),]
unrate_monthly = head(unrate_monthly, -16)

ffrate_monthly = ffrate_monthly[-(1:78),]
ffrate_monthly = head(ffrate_monthly, -16)

recession = recession[-(1:1274),]
recession = head(recession, -17)
# Convert monthly data to annual by averaging each years values
unrate_annual = vector()
ffrate_annual = vector()
total_unrate = 0
total_ffrate = 0

for (x in 1:744) {
  total_unrate = total_unrate + unrate_monthly$UNRATE[x]
  total_ffrate = total_ffrate + ffrate_monthly$FEDFUNDS[x]

  if (x%%12 == 0){
    average_unrate = total_unrate/12
    average_ffrate = total_ffrate/12
    unrate_annual = append(unrate_annual, average_unrate)
    ffrate_annual = append(ffrate_annual, average_ffrate)
    total_unrate = 0
    total_ffrate = 0
  }
}
# Determine "category" of year based on if it occurs in one of its months
library("lubridate")

recession$DATE = as.Date(recession$DATE)
recession$Year = year(recession$DATE)
```

```r
recession_new = data.frame(year = unique(recession$Year),
                           recession = ifelse(unique(recession$Year)
                                  %in% recession$Year[recession[2] == 1], 1, 0))
# Create dataframe out of cleaned data
financial_data = data.frame(date = gdpgrowth$Date, returns = returns$value,
                            ffrate = ffrate_annual, unrate = unrate_annual,
                            inflation = inflation$Inflation.Rate,
                            gdpgrowth = gdpgrowth$GDP.Growth,
                            prioryear_return = returns_prioryear$value,
                            recession = recession_new$recession
)
# Create dateless version for correlation matrix,
  # quantitative variables only version for pairs plot
dateless = subset(financial_data, select = -c(date))
quant_only = subset(financial_data, select = -c(date, recession))
# Pairs plot to examine relationship between quantitative variables
pairs(quant_only)
library("kableExtra")

kable(table(financial_data$recession),
      col.names = c("Value (1 = Recession Year)", "Count")) %>%
  kable_styling()
# Correlation matrix with correlations and p-values
library("Hmisc")
rcorr(as.matrix(dateless), type = c("pearson","spearman"))
model_all <- lm(returns ~ inflation + ffrate + unrate + prioryear_return + recession,
                data = financial_data)
summary(model_all)
model_4var <- lm(returns ~ inflation + unrate + prioryear_return + recession,
                 data = financial_data)
model_3var <- lm(returns ~ inflation + unrate + recession, data = financial_data)
model_2var <- lm(returns ~ unrate + recession, data = financial_data)
model_1var <- lm(returns ~ unrate, data = financial_data)
predictornames = c("inflation + ffrate + unrate + prioryear_return + recession",
               "inflation + unrate + prioryear_return + recession",
               "inflation + unrate + recession", "unrate + recession",
               "unrate")
rsquared = c(summary(model_all)$r.squared, summary(model_4var)$r.squared,
             summary(model_3var)$r.squared, summary(model_2var)$r.squared,
             summary(model_1var)$r.squared)
adjrsquared = c(summary(model_all)$adj.r.squared, summary(model_4var)$adj.r.squared,
             summary(model_3var)$adj.r.squared, summary(model_2var)$adj.r.squared,
             summary(model_1var)$adj.r.squared)

modelselectdf = data.frame(predictornames, rsquared, adjrsquared)

kable(modelselectdf, col.names = c("Predictors Used in Model", "R-Squared",
                                   "Adjusted R-Squared")) %>%
  kable_styling()
cross_valid = function(model, k = nrow(financial_data), data = financial_data) {
  n = nrow(data)
  res = 0
  index = matrix(1:n, ncol = k)
```

```
  for (i in 1:k) {
    train = data[-index[, i], ]
    test = data[index[, i], ]
    if (model == 1) {
      lm_fit=lm(returns ~ inflation + ffrate + unrate + prioryear_return + recession, data=train)
    } else if (model == 2) {
      lm_fit=lm(returns ~ inflation + unrate + prioryear_return + recession, data=train)
    } else if (model == 3) {
      lm_fit=lm(returns ~ inflation + unrate + recession, data=train)
    } else if (model == 4) {
      lm_fit=lm(returns ~ unrate + recession, data=train)
    } else if (model == 5) {
      lm_fit=lm(returns ~ unrate, data=train)
    }
    y_pred=predict(lm_fit, test)
    res=res + mean((y_pred - test$returns)^2)
  }
  res = res / k
  return(res)
}
set.seed(1523)
sampledata = financial_data[sample(1:nrow(financial_data)), ]

fivkfold = c(cross_valid(1, 5, sampledata), cross_valid(2, 5, sampledata),
             cross_valid(3, 5, sampledata), cross_valid(4, 5, sampledata),
             cross_valid(5, 5, sampledata))
tenkfold = c(cross_valid(1, 10, sampledata), cross_valid(2, 10, sampledata),
             cross_valid(3, 10, sampledata), cross_valid(4, 10, sampledata),
             cross_valid(5, 10, sampledata))
loocv = c(cross_valid(1), cross_valid(2), cross_valid(3), cross_valid(4), cross_valid(5))

crossvaldf = data.frame(predictornames, fivkfold, tenkfold, loocv)

kable(crossvaldf, col.names = c("Predictors Used in Model", "K=5-fold",
                                "K=10-fold", "LOOCV")) %>%
  kable_styling()
library("MPV")

PRESS <- function(model) {
  residuals <- residuals(model)
  hat_values <- hatvalues(model)
  press <- sum((residuals / (1 - hat_values))^2)
  return(press)
}
All.Criteria = function(the.model){
  p =length(the.model$coefficients)
  n =length(the.model$residuals)
  the.LL = logLik(the.model)
  the.BIC =-2*the.LL + log(n)*p
  the.AIC =-2*the.LL + 2*p
  the.PRESS = PRESS(the.model)
  the.R2adj = summary(the.model)$adj.r.squared
  the.results = c(p, the.LL, the.AIC, the.BIC, the.PRESS, the.R2adj)
```

```
  names(the.results) = c("p", "LL","AIC","BIC","PRESS", "R2adj")
  return(the.results)
}

modeldata<- data.frame(y=financial_data$returns,x1=financial_data$unrate,x2=financial_data$ffrate, x3=fi
full.model<- lm(y~x1+x2+x3+x4+x5, data = modeldata)
All.Models<- c("y~x1+x2+x3+x4+x5","y~x1+x3+x4+x5", "y~x1+x3+x5", "y~x1+x5","y~x1")
all.model.crit = t(sapply(All.Models, function(M){
  current.model = lm(M, data = modeldata)
  All.Criteria(current.model)
 }))
all.model.crit
library("ggplot2")

ggplot(data = financial_data,
       mapping = aes(x = unrate + recession,
                     y = returns)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Annual S&P 500 Returns vs Predictors",
       x = "Unemployment Rate + Recession Year",
       y = "S&P 500 Returns")
final.model<- lm(returns~unrate+recession, data = financial_data)
par(mfrow=c(2,2))
plot(final.model)
fix.model<-lm(returns^2~unrate+recession, data = financial_data)
par(mfrow=c(2,2))
plot(fix.model)
```