**Predict Stock Market Crash**

Group 6: Fuyu Liu, Zhuoping Fan, Ziqian Ding

**1 Introduction**

This report presents a focused regression analysis to identify key factors influencing S&P 500's monthly returns in order to study the model's ability to predict market crashes. Starting with a simple linear regression, we isolate the impact of individual independent variables on market behavior. Moving on to multiple linear regressions, we fully assess the combined impact of all variables. Emphasis is placed on model selection for accurate market return predictions, incorporating dummy and logit regressions for insights into market volatility and bear trends. This coherent analysis aims to provide a holistic understanding of S&P500 dynamics, offering valuable insights for stakeholders to navigate and interpret stock market movements.

**2 Data Preparation**

**2.1 Data Source**

The dataset for this study is sourced from the Federal Reserve Economic Data (FRED), a comprehensive database maintained by the Federal Reserve Bank of St. Louis. FRED offers a wide range of economic data, which is crucial for conducting in-depth financial analyses and developing robust predictive models.

**2.2 Selection of Variables**

To construct a predictive model for stock market crashes, we have selected an indicator from each category provided by FRED. These economic indicators are selected as regression variables due to their potential influence on market dynamics:

* *Federal Funds Effective Rate:* The federal funds rate is the interest rate at which depository institutions trade federal funds with each other overnight. It is the central interest rate in the U.S. financial market. It is extremely informative about future movements of real macroeconomic variables (Ben, 1990) such as the prime rate, mortgages, loans, and savings, all of which are very important to consumer wealth and confidence.
* *Total Federal Outlays:* Federal spending policies, represented by total federal outlay, can affect the economy through their impacts on federal borrowing, private demand for goods and services, people's incentives to work and save, federal investment, etc, providing implications for the stock market.
* *House Price Index:* The housing market is a significant component of the overall economy. House price index is a broad measure of the movement of single-family house prices, fluctuations in this index can indicate broader economic shifts that affect stock markets. (Federal Housing Finance Agency, 2023)
* *Unemployment Rate:* The unemployment rate represents the number of unemployed as a percentage of the labor force, which is a key indicator of economic health and consumer sentiment. Change in this rate has a strong impact on stock prices (Gonzalo & Taamouti, 2017).
* *Producer Price Index (PPI) for all commodities:* PPI measures the average change in prices received by producers for their goods and services. It is considered a leading indicator as it reflects changes in production costs. Mishkin (1982) argues that significant changes in producer prices may signal disruptions in supply chains, impacting corporate profitability and potentially foreshadowing economic downturns and market crashes.
* *Consumer Price Index for all items in Northeast (CPI):* CPI gauges the average change in prices paid by urban consumers for a basket of goods and services. It provides insights into inflationary pressures. Perner et al. (1998) suggest that deviations in CPI from expectations can be indicative of economic stress and may precede market downturns.
* *Real-time Sahm Rule Recession Indicator:* Sahm's Rule is a real-time recession indicator based on the national unemployment rate, providing timely insights into economic conditions. Sahm (2019) demonstrates the effectiveness of this rule in predicting recessions, which are closely tied to market crashes.
* *Consumer Opinion Surveys: Confidence Indicators: Composite Indicators: OECD Indicator for the United States:* Composite confidence indicators, such as the OECD Indicator, reflect consumer and business sentiment, offering a comprehensive view of economic expectations. Barsky et al. (2009) argue that declining consumer and business confidence may precede reduced spending and investment, contributing to economic downturns and market crashes.

**2.3 Data Collection**

The above data, covering years from 1990 to 2022, was downloaded in a CSV format and titled 'P\_LIU\_FAN\_DING.csv'. This time range was chosen to capture a variety of economic conditions, including growth periods, recessions, and recoveries, to enhance the predictive model's accuracy and robustness.

**3 Modeling Process and Results**

In this section, we conducted simple linear regression, multiple linear regression, model selection, dummy regression and logit regression to build the best model for prediction. It should be noted that we use 95% confidence intervals by default for coefficient estimates.

**3.1 Simple Linear Regression**

Simple linear regression allows for an initial understanding of each variable's individual impact. In this section, we defined each variable and then used Single Linear Regression to regress Monthly Price Return with 8 variables separately. The results are shown in Figure 1.

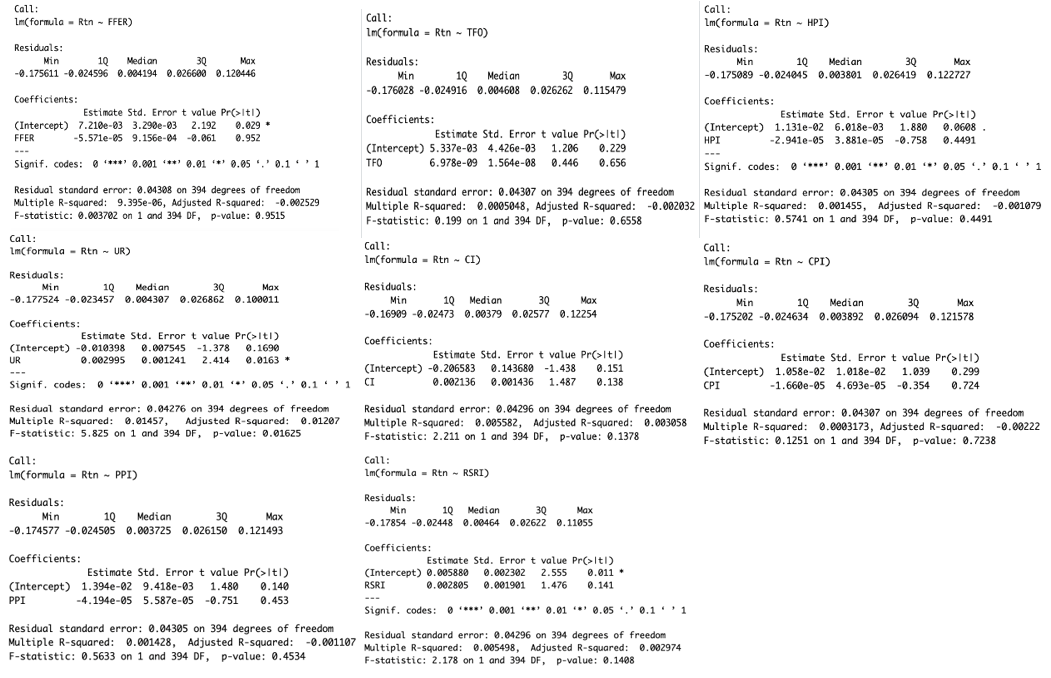


Figure 1: Results of Simple Linear Regression

Based on the regression results of monthly price return with eight explanatory variables, we find that the p-value for the Unemployment Rate (UR) is less than 0.05, which means that this variable is significant and meaningful. On the contrary, the p-values of the other variables are greater than 0.05, which means that these variables do not have a significant impact on the prediction results based on this regression.

**3.2 Multiple Linear Regression**

In this section, we use Multiple Linear Regression to test the overall relationship between Monthly Price Return and eight other variables. Results are shown in Figure 2.

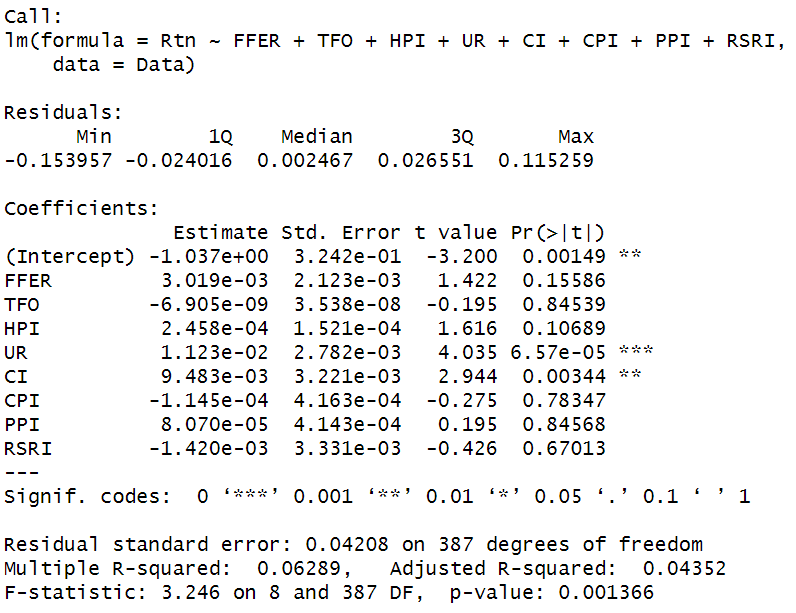


Figure 2: Results of Multiple Linear Regression

From the result of the multiple linear regression, we can find that only the P-value of the variable of Confidence Indicators (CI) is less than 0.05, which means that this variable is significant for predicting. And the P-value of the other variables is greater than 0.05, which means that those variables are not meaningful for the prediction results based on this regression.

**3.3 Model Selection**

In the process of model selection to identify the best multiple regression model for predicting the S&P500 monthly return, we began with eight independent variables and one dependent variable. Employing multiple regression in R, we utilized subset selection, Stepwise AIC and Stepwise Cp methods to refine our model.

**Subset Selection:**  
Subset selection involves systematically evaluating all possible combinations of predictor variables to determine the subset that optimally predicts the dependent variable. Figure 3 illustrates the results of subset selection, showcasing the variables selected for different subset sizes.

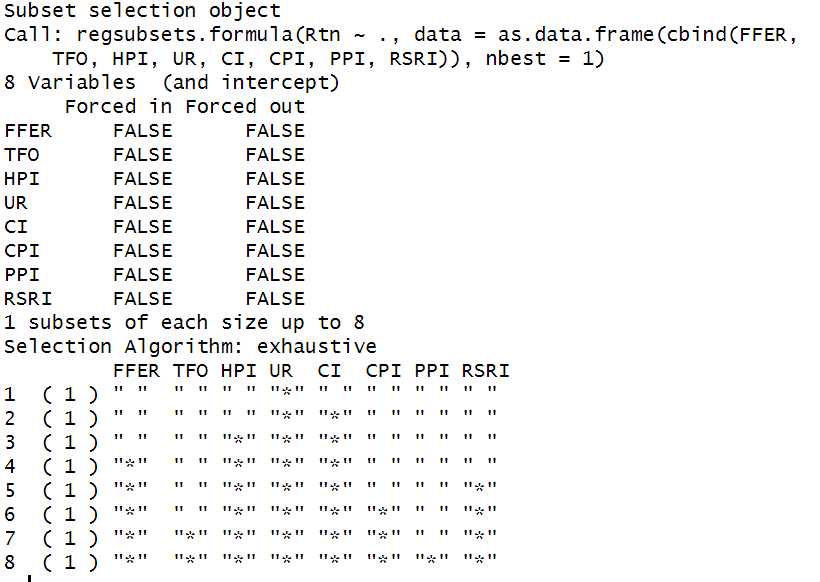


Figure 3: Results of Subset Selection

**Stepwise AIC:**  
We proceeded with Stepwise AIC, a backward variable selection method aimed at minimizing the AIC. The logic behind Stepwise AIC is to iteratively remove variables to improve the model fit. Figure 4 displays the result, revealing that the final model includes four variables: The Federal Funds Effective Rate (FFER), House Price Index (HPI), Unemployment Rate (UR), and Consumer Confidence Index (CI), with the minimum AIC of -2507.67.

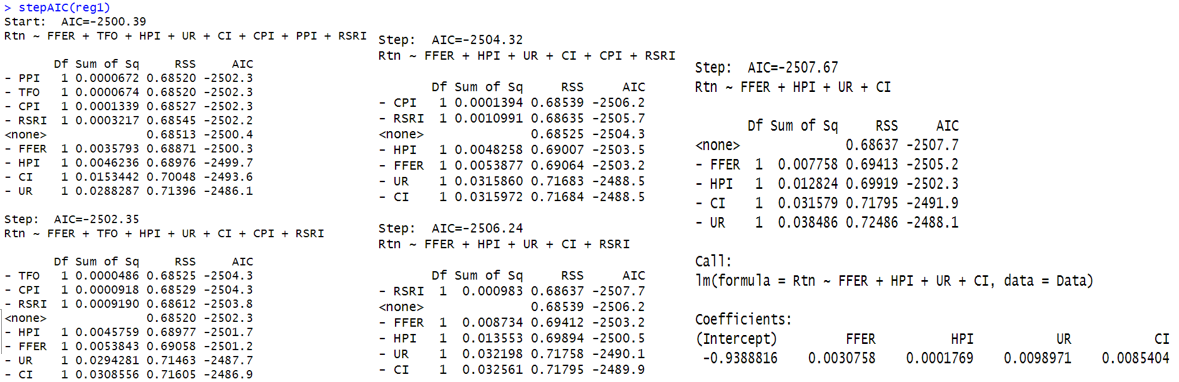


Figure 4: Results of Stepwise AIC

**Stepwise Cp:**  
To validate the robustness of our model selection, we employed Stepwise Cp, which also aims to identify the optimal subset of variables. Figure 5 indicates that, consistent with Stepwise AIC, the final model includes the same four variables, demonstrating the reliability of our selection.

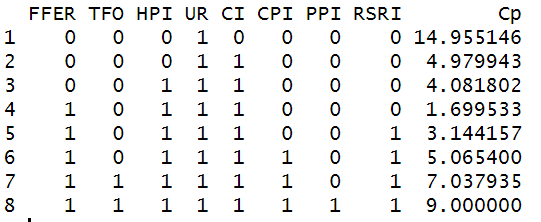


Figure 5: Results of Stepwise Cp

Subsequently, we conducted multiple regression again using the chosen variables (FFER, HPI, UR, and CI). Figure 6 summarizes the regression results. Notably, the p-values for FFER (0.03617), HPI (0.00718), UR (3.92e-06), and CI (2.78e-05) are all statistically significant. The positive coefficients suggest that an increase in these variables is associated with an increase in the S&P500 monthly return.

However, the model's R-squared value of 0.0612 indicates that only 6.12% of the variability in the S&P500 return is explained by these four variables. This implies that there are other relevant factors not considered in our model, prompting further exploration.

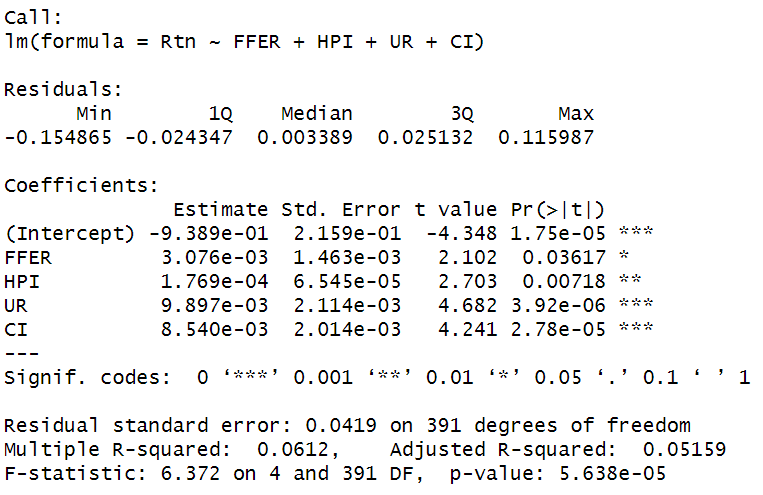


Figure 6: Results of Selected Regression Model

In conclusion, the consistency in variable selection across Subset Selection, Stepwise AIC, and Stepwise Cp reinforces the credibility of our chosen model. While the identified variables exhibit significance, the modest R-squared underscores the need for additional variables to enhance the predictive power of our model.

**3.4 Dummy Regression**

In this section, we try to robust the model by adding dummy variables and testing seasonal effects on the stock market. It is generally believed that August, September, and October are risky months for stock markets, thus we incorporated dummy variables corresponding to these three months. Together with the four chosen variables in Section 3.3, the summary of the regression is shown in Figure 7.

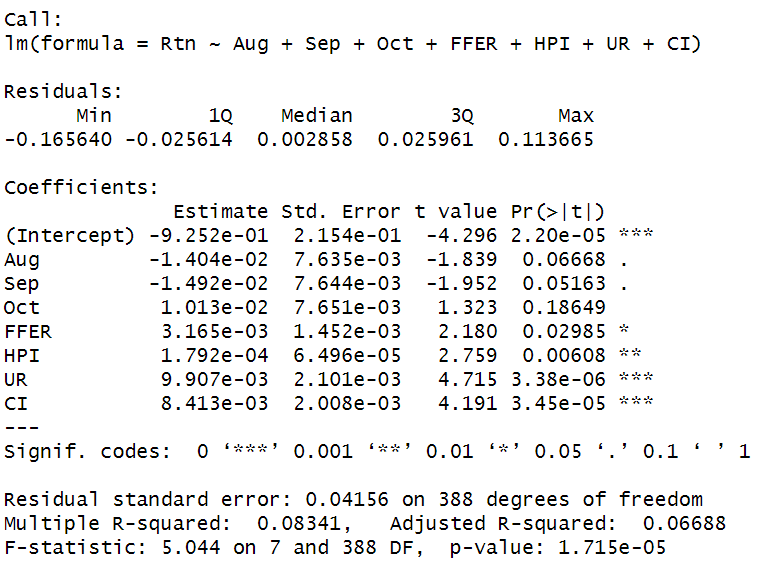


Figure 7: Results of Dummy Regression

The dummy regression model's summary indicates that among the variables considered, the Unemployment Rate (UR) and the Consumer Confidence Index (CI) have the most significant impact on the stock market return (Rtn), with both showing strong statistical significance. The Federal Funds Effective Rate (FFER) and House Price Index (HPI) also demonstrate a statistically significant relationship with Rtn, but in a weaker way. The monthly dummy variables (August, September, October) do not show a statistically significant impact on Rtn. While the coefficients for August and September are negative and that for October is positive, their p-values suggest these relationships are not statistically robust.

The model’s adjusted R-squared value of 0.06688 indicates that it explains approximately 6.7% of the variance in Rtn, which is relatively low, suggesting that other unaccounted factors might also be influencing Rtn. The F-statistic is significant, indicating that the model is statistically significant overall compared to a model with no predictors.

In summary, while the economic indicators (UR, CI, FFER, HPI) have a notable influence on stock returns, the three specific months within the year do not show a significant impact on this model. Also, there may be other unforeseen factors that can better explain the Rtn.

**3.5 Logit Regression**

Logit regression is a statistical method used for binary classification problems, particularly when the dependent variable is categorical. In the context of predicting bear markets, where the criterion is a monthly stock market decline of at least 4%, logit regression helps model the probability of the event occurring. The logistic function is employed to constrain the output between 0 and 1, representing the probability of a bear market.

**3.5.1 Selection of Data and Logit Regression**  
In the initial step, returns below -4% were identified as instances of bear markets. Subsequently, a logit regression was performed using the selected independent variables from the optimal multiple regression model identified in the earlier step. Coeftest, as shown in Figure 8, was employed to conduct z-tests of the coefficients, providing insights into the significance of each variable in predicting bear markets.

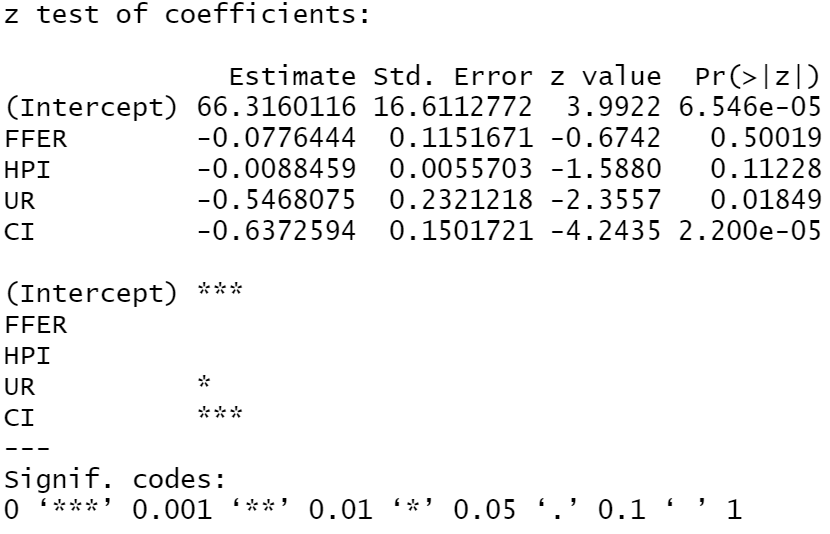


Figure 8: Results of Coeftest

**3.5.2 Interpreting Coeftest Results**

The coeftest results, as displayed in Figure 8, provide insights into the significance of each coefficient in predicting bear markets. The intercept, representing the baseline probability of a bear market, has a p-value of 6.546e-05, indicating its statistical significance. Among the selected variables, the Unemployment Rate (UR) and Consumer Confidence Index (CI) also exhibit significance with p-values of 0.01849 and 2.200e-05, respectively. Notably, the Intercept and CI have particularly high levels of significance.

All coefficients in the logit regression are negative, implying that a decrease in these variables is associated with an increased likelihood of entering a bear market. This aligns with the expectation that economic indicators, such as unemployment and consumer confidence, declining could precede a market downturn.

In conclusion, the logit regression results suggest that the selected independent variables, including the intercept, Unemployment Rate, and Consumer Confidence Index, play a significant role in predicting bear markets defined by a monthly decline of 4% or more. The high significance levels of the intercept and CI underscore their importance in determining the likelihood of entering a bear market. However, the overall model performance should be considered in conjunction with additional metrics, and further exploration may enhance the predictive power of the model.

**4 Conclusion**

In this comprehensive analysis, we utilized data sourced from FRED to construct a predictive model for stock market crashes, considering various economic indicators. Single linear regression highlighted unemployment rate as a significant variable, while multiple linear regression showed the confidence indicator as a key predictor for monthly stock returns. Our model selection processes, including subset selection, Stepwise AIC, and Stepwise Cp, consistently identified four critical variables (Federal Funds Effective Rate, House Price Index, Unemployment Rate, and Consumer Confidence Index) as the optimal combination for predicting the S&P500 monthly return.

However, the modest R-squared suggested the need for additional variables. Incorporating dummy variables for perceived risky months and conducting logit regression to predict bear markets revealed that Unemployment Rate and Consumer Confidence Index significantly influenced the likelihood of a bear market. While the model's overall significance was confirmed, the need for further exploration and consideration of additional factors to enhance predictive power was emphasized. The findings showcase the complexity of stock market prediction and underscore the ongoing pursuit of a more comprehensive understanding of contributing factors.

**Reference**

1. Barsky, Robert B. and Sims, Eric R.(2009, June) *Information, Animal Spirits, and the Meaning of Innovations in Consumer Confidence*. NBER Working Paper No. w15049. <https://ssrn.com/abstract=1415212>
2. Bernanke, B. (1990, October 1). *The federal funds rate and the channels of monetary transmission*. NBER. https://www.nber.org/papers/w3487
3. Federal Housing Finance Agency. (2023). *FHFA House Price Index.* https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index.aspx
4. FRED. (2023, November 30). *Federal funds effective rate*. https://fred.stlouisfed.org/series/DFF
5. Gonzalo, J. & Taamouti, A. (2017). The reaction of stock market returns to unemployment. *Studies in Nonlinear Dynamics & Econometrics*, *21*(4), 20150078. <https://doi.org/10.1515/snde-2015-0078>
6. Mishkin, F. S. (1982). Does Anticipated Monetary Policy Matter? An Econometric Investigation. *Journal of Political Economy*, *90*(1), 22–51. http://www.jstor.org/stable/1831228
7. Perner, L. (1998). Asking About Prices: A New Approach to Understanding Price Stickiness. *The Journal of Consumer Affairs, 32*(2), 424-427. https://www.proquest.com/scholarly-journals/asking-about-prices-new-approach-understanding/docview/195907066/se-2
8. Sahm, C. R. (2019). Recession Ready. *The Hamilton Project. https://www.hamiltonproject.org/publication/policy-book/recession-ready-fiscal-policies-to-stabilize-the-american-economy/*
9. U.S. Bureau of Labor Statistics. (November 24, 2023). FRED. *Unemployment Rate [UNRATE].* https://fred.stlouisfed.org/series/UNRATE