REGRESSION ANALYSIS PROJECT

Regression Analysis of US Bank Stock Returns with Macroeconomic and Peer Bank Effects

Submitted by – Group 5

Members	**Roll Number**
Kaustab Saha	24N0073
Santanu Biswas	24N0072
Aniruddha Mukherjee	24N0075
Sourasish Saha	24N0077
Priom Pal	24N0079
Sourav Mahara	24N0080

Supervisor: Prof. Monika Bhattacharjee

Department of Mathematics

INDIAN INSTITUTE OF TECHNOLOGY BOMBAY

Contents

1 Statement Of The Problem

> *"The four most dangerous words in investing are: 'this time it's different.'"*
> — **Sir John Templeton**

The stability of the financial sector is closely tied to the performance of major banks, which serve as the backbone of any economy. Over the past two decades, global financial markets have faced several major disruptions. These economic disasters have significantly impacted investor behavior, market stability and the financial sector's performance.

This project explores the stock return patterns of leading American banks during this turbulent time frame. Our primary goal is to uncover how external macroeconomic factors and inter-bank relationships influence return behavior across these financial institutions.

The central challenge of this project lies in building a robust **multiple linear regression model** that captures weekly log returns of these banks amidst periods of economic volatility. By aligning historical stock performance with macroeconomic indicators, we seek to provide deeper insights into the resilience, interconnectedness and vulnerabilities of the banking sector — ultimately supporting more informed decision-making in financial risk management and policy formulation.

2 Selection of Potential Variables

To analyze the evolving dynamics of the financial sector, we consider weekly stock return data from a range of financial institutions—including commercial banks, investment banks, and insurance firms—across a long time span. This diverse dataset helps capture both sector-specific behaviors and systemic market trends.

To better understand external influences, we incorporate several macroeconomic indicators known to impact financial markets:

- **Inflation Rate (CPI):** Rising inflation erodes consumer purchasing power and often prompts tighter monetary policy, which can dampen economic growth and affect bank profitability.
- **Gross Domestic Product (GDP):** As a measure of overall economic activity, GDP growth indicates stronger demand for financial services and lower default risk.

- **S&P 500 Index:** Acts as a barometer of investor sentiment and overall market performance, affecting capital flows and financial sector valuations.
- **Federal Funds Rate:** The interest rate set by the Federal Reserve influences borrowing costs, liquidity, and the yield curve—key drivers of bank income and lending behavior.

To account for structural breaks and external shocks, we incorporate impacts for major economic and financial events. These include the **9/11 Attacks** (September 2001), which triggered market panic and geopolitical uncertainty; the **Enron Bankruptcy** (December 2001), raising concerns over corporate transparency; the **Lehman Brothers Collapse** (September 2008), which marked the peak of the global financial crisis; the **TARP Bailout** (October 2008), aimed at stabilizing the banking system; the **Federal Reserve's QE1** (November 2008), initiating large-scale monetary stimulus; the **Dodd-Frank Act** (July 2010), introducing significant regulatory reforms; the **US Debt Downgrade** by S&P (August 2011), affecting investor sentiment; the **Taper Tantrum** (May 2013), which led to heightened market volatility over tightening monetary policy; and the **Brexit Vote** (June 2016), which caused global uncertainty with potential implications for internationally exposed financial firms.

This combination of firm-level data, macroeconomic indicators, and high-impact event controls allows for a comprehensive understanding of the forces influencing bank stock returns.

3 Data Collection

We collected stock price data for the top six U.S. banks by total assets—*JPMorgan Chase, Bank of America, Citigroup, Wells Fargo, Goldman Sachs*, and *Morgan Stanley*. The dataset spans from **January 2, 2001 to December 27, 2016**, with data recorded at a **weekly frequency**, aligning with our model's temporal resolution.

In addition to firm-level data, we incorporated relevant **macroeconomic indicators** to better capture the broader economic environment. Specifically, we gathered:

- **Gross Domestic Product (GDP)** on a *quarterly basis*, sourced from the *Federal Reserve Economic Data (FRED)* database.
- **Federal Funds Rate** on a *daily basis*, sourced from *FRED*.

Rank	Bank	HQ	Total assets (billions of US$)[3]
1	JPMorgan Chase	New York City	$4,210
2	Bank of America	Charlotte, North Carolina	$3,324
3	Citigroup	New York City	$2,430
4	Wells Fargo	San Francisco, California	$1,922
5	Goldman Sachs	New York City	$1,728
6	Morgan Stanley	New York City	$1,258

Top 5 U.S. Banks by Total Assets *(as of 2025)*

- **S&P 500 Index** on a *weekly basis*, obtained from *Yahoo Finance*.
- **Inflation Rate (CPI-based)** on a *monthly basis*, also sourced from *FRED*.

All macroeconomic series were aligned to the weekly stock return data using cubic spline technique, ensuring consistency across the dataset.

Note: All data used in the study were publicly available and retrieved from reliable sources such as the Federal Reserve Economic Data (FRED) platform and Yahoo Finance.

4 Exploratory Data Analysis and Data Preprocessing

4.1 Checking for Missing Values

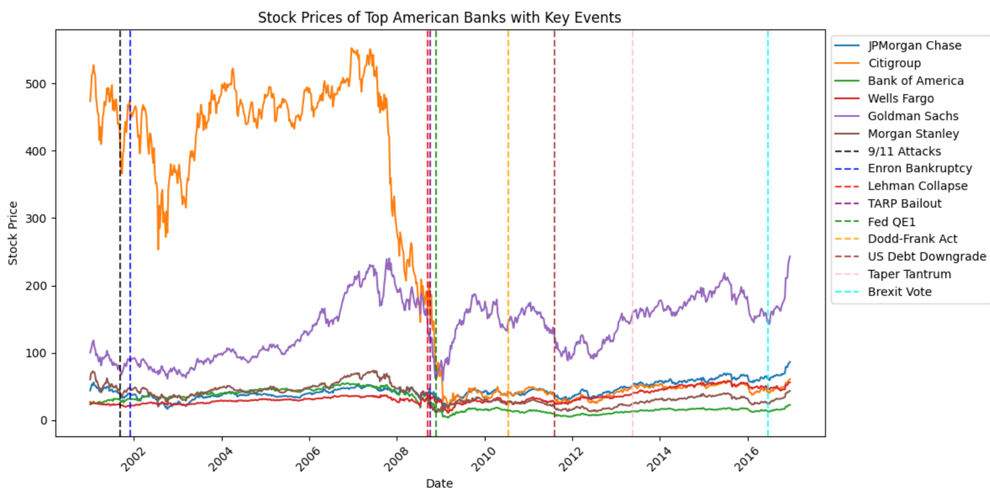
As a preliminary step, we checked for missing values across all variables in the data set and confirmed that there were no null entries.

4.2 Dataset Overview

The dataset contains 824 observations and 11 variables, including the date column, stock returns of six major U.S. banks, the S&P 500 index, and three key macroeconomic indicators. To gain deeper insights into the behavior of bank stock performance and macroeconomic indicators over time, we visualized their time series in conjunction with significant economic and financial events.

4.3 Bank Stock Prices with key events

The plot below illustrates the stock prices of six leading US banks over time. Key economic events such as the 9/11 attacks, Lehman Brothers collapse, and TARP bailout are marked with vertical lines.

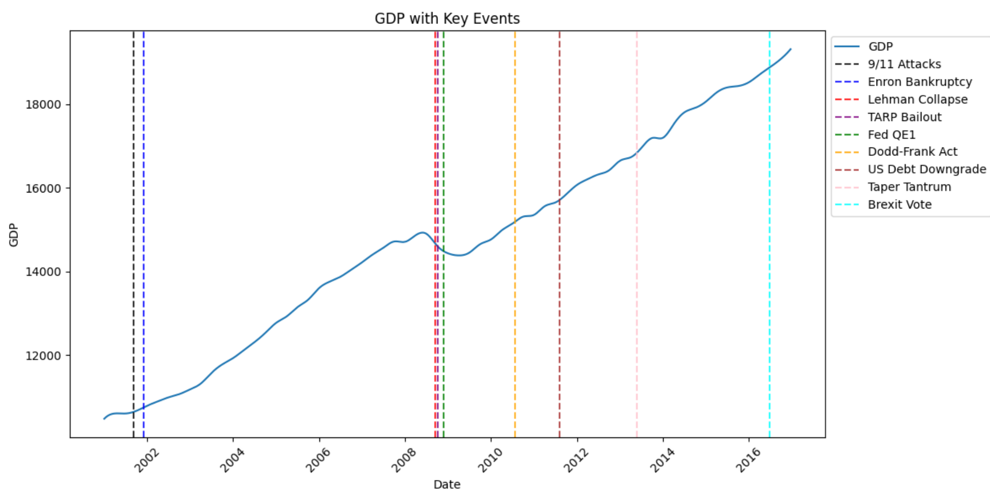


Citigroup experienced a dramatic decline during the 2008 financial crisis, emphasizing its vulnerability. Other banks also show visible downturns, particularly in response to crisis events, followed by slow recoveries.

Post-2010, stock prices generally trended upward, reflecting stabilization efforts and economic recovery. While the time series visualization of stock prices provides an initial understanding of the performance trends across different banks, it has notable limitations. The absolute price levels of the stocks vary significantly, making direct comparisons less meaningful. Additionally, corporate actions such as stock splits and differences in price scales can distort the interpretation of volatility and relative performance. To address these inefficiencies and enable a more standardized comparison, we shift our focus to stock returns, which normalize price changes and better capture the underlying financial dynamics across banks.

4.4 GDP with Key Events

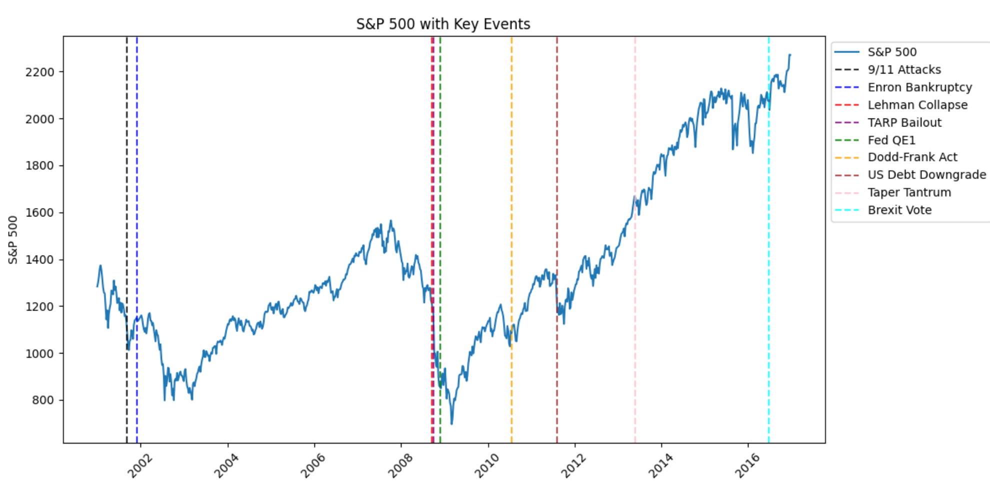
The plot below shows the US GDP trend over time, again with key economic events overlaid:



GDP showed consistent growth until 2008, where a brief decline is visible during the financial crisis. Following the Federal Reserve's QE1 and other stimulus measures, GDP recovered steadily, highlighting the impact of macroeconomic policy interventions.

4.5 S&P 500 Index with Key Events

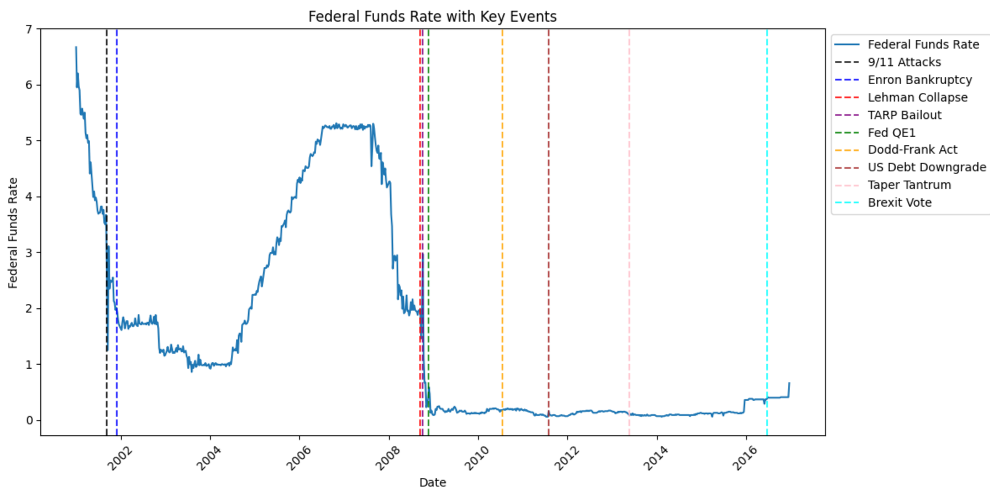
The graph shows the S&P 500 index along with vertical lines marking significant economic or political events.



Post-2008 financial crisis, the S&P 500 dropped drastically and rebounded slowly after interventions like the Federal Reserve's QE1. Events like the US Debt Downgrade and Brexit Vote show minor dips and recoveries.

4.6 Federal Funds Rate with Key Events

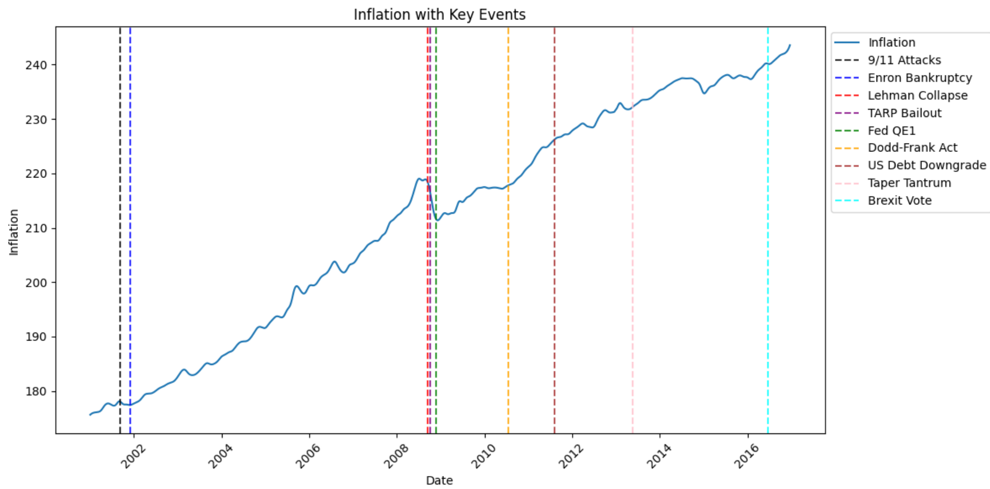
The plot below shows Federal Funds Rate change with major economic events.



Significant decrease around the 2008 crisis and remained low until a gradual rise post-2015. Key events like QE1 and the TARP bailout coincide with near-zero rates indicating expansionary monetary policy

4.7 Inflation with Key Events

The below plot tracks inflation from 2001 to 2016 alongside major economic events.



The Figure shows steady increase until the 2008 financial crisis. Slight dip post-Lehman collapse and TARP, followed by a continued upward trend.

4.8 Stationarity Check

Before performing regression analysis, it is essential to check the stationarity of the data, especially for time series models. We conducted an Augmented Dickey-Fuller (ADF) test on all variables to test for stationarity. The results indicated that none of the 10 variables were stationary. Since stationarity is a key assumption for time series models, and to address the issue of significant variation in absolute price levels across the stock data, we applied necessary transformations. These transformations ensure that the data is stationary and suitable for further analysis.

4.8.1 Transformation of Variables

Initially, we considered two types of transformations for the data: returns and log-returns. However, we discarded the log-return transformation due to the nature of the banking sector, which tends to exhibit stable and relatively consistent returns. Since the returns are generally close to 1, applying a log transformation would result in values close to zero, reducing the variation in the data. This lack of variation would make it difficult to distinguish between values and complicate the estimation process. Consequently, we opted to use returns for our analysis, as this transformation preserves the necessary variation and allows for more effective modeling. The transformation we used is

$$Returns_t = Price_t / Price_{t-1} - 1$$

where,
$Price_t$ indicates the price at time point t and $Returns_t$ indicates return at time point t.For the banks we used this transformation and for the macroeconomic factors we used differencing. The dataset is now stationary, as confirmed by the ADF test, making it suitable for subsequent regression analysis.

4.9 Autocorrelation Function

To better understand the temporal structure of each variable, we examined their Autocorrelation Function (ACF) plots. These plots illustrate how each time series correlates with its own past values (lags). Spikes outside the 95% confidence interval indicate statistically significant autocorrelations.

4.9.1 Bank Stock Returns (JPMorgan Chase, Citigroup, Bank of America, Wells Fargo, Goldman Sachs, Morgan Stanley)

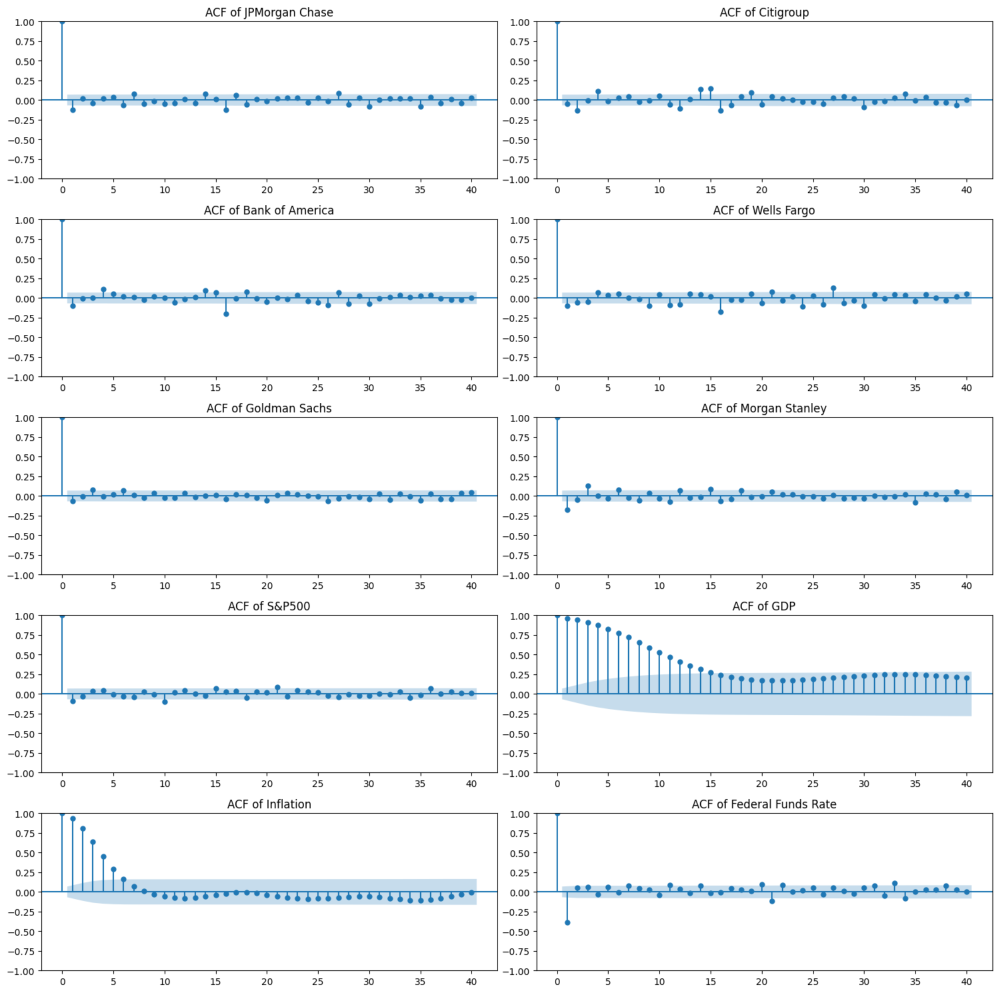
The ACF plots for all six bank stocks reveal low or insignificant autocorrelation beyond lag 0. This implies that the weekly stock returns are weakly autocorrelated, which is typical for financial return series that often behave like white noise.

4.9.2 S&P 500 Return

The S&P 500 return series also shows minimal autocorrelation, reinforcing the idea that market returns are largely unpredictable. This is consistent with the Efficient Market Hypothesis, which suggests that past values cannot reliably predict future returns.

4.9.3 GDP

The ACF plot for GDP exhibits a smooth, slow decay, remaining significantly positive across multiple lags. This is characteristic of a non-stationary or trend-following process. Such a pattern suggests that the GDP series needs differencing before it can be used effectively in regression analysis.



4.9.4 Inflation

Inflation shows significant autocorrelation up to around lag 8, with a gradual decline afterward. This indicates a persistent and slowly adjusting process, possibly non-stationary or mean-reverting in nature. Appropriate transformations, such as differencing, may be required depending on the modeling approach.

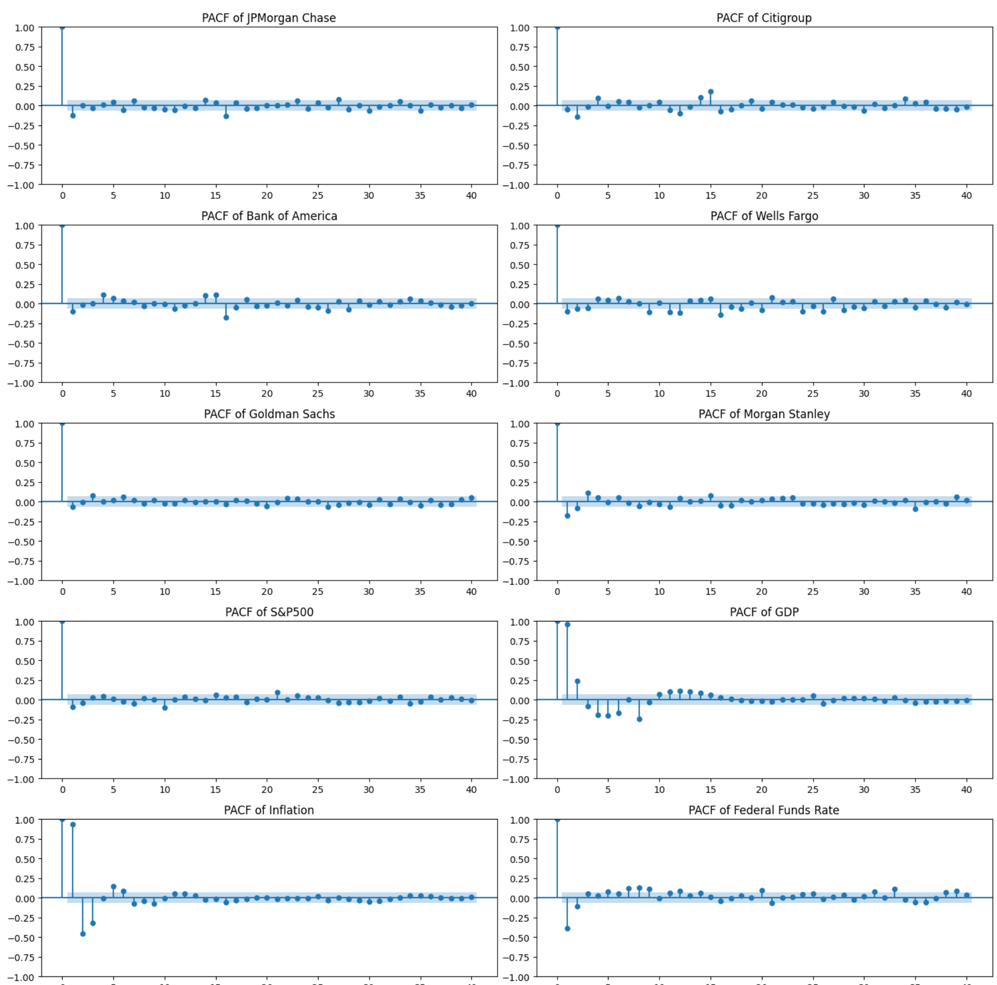
4.9.5 Federal Funds Rate

The ACF of the Federal Funds Rate shows low but significant autocorrelation at short lags, indicating the presence of short-term memory in the data. This may justify the inclusion of lagged values in regression or time series forecasting models.

So from the above figures we can summarize Stock returns and S&P 500: Show weak or no autocorrelation, suggesting they may already be stationary. GDP, Inflation, and Federal Funds Rate: Exhibit significant autocorrelation, indicating potential non-stationarity. These variables may require differencing or lag inclusion before regression analysis. Care must be taken to avoid spurious regression when combining non-stationary variables (e.g., GDP) with stationary ones (e.g., returns).

4.10 Partial Auto Correlation Function

The PACF measures the direct relationship between a time series and its lagged values by filtering out the effects of shorter lags. This shortens our perspective on which lags are most influential, aiding in model specification, particularly for autoregressive models. Below is the figure displaying the PACF analysis along with the key observations drawn from it



4.10.1 Bank Stock Returns (JPMorgan Chase, Citigroup, Bank of America, Wells Fargo, Goldman Sachs, Morgan Stanley)

Each ACF plot shows minimal to no significant autocorrelation beyond lag 0.Weekly stock returns behave similarly to white noise, suggesting little predictability from past returns.

4.10.2 S&P 500

ACF plot reveals weak or statistically insignificant autocorrelation across lags. In line with the Efficient Market Hypothesis, indicating that past index returns have limited predictive power for future returns.

4.10.3 GDP

The ACF decays slowly and remains positive for multiple lags, indicating persistent or trend-like behavior. Non-stationary series; likely requires differencing before regression to avoid spurious results.

4.10.4 Inflation

Significant autocorrelation up to around lag 8, with a gradual decline afterward. Potentially non-stationary or mean-reverting. May benefit from differencing, depending on model specifications.

4.10.5 Federal Funds Rate

Shows low but statistically significant autocorrelation at shorter lags.Short-term memory in monetary policy rates, suggesting lagged values can be relevant predictors in time series models.

5 Model Specification

5.1 Definition

Linear regression assumes a linear relationship between the independent variables (also known as predictor variables or features) and the dependent variable (also known as the response variable). It seeks to model this relationship with a linear equation.

5.2 Equation of the Model

The general form of a multiple linear regression model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i \tag{1}$$

Where:

- Y_i is the value of the dependent variable for the i-th observation.
- $X_{1i}, X_{2i}, \ldots, X_{pi}$ are the values of the p independent variables for the i-th observation.
- β_0 is the intercept (the value of Y when all X variables are zero).
- $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients (slopes) that represent the change in Y for a one-unit change in the corresponding X_i variable, holding all other X_j (where $j \neq i$)) variables constant.
- ϵ_i is the random error term for the i-th observation.

In matrix form, the linear regression model can be expressed as:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E} \tag{2}$$

where:

- $\mathbf{Y}$ is an $n \times 1$ vector of the observed values of the dependent variable.
- $\mathbf{X}$ is an $n \times (p+1)$ matrix of the observed values of the independent variables (including a column of 1s for the intercept).
- β is a $(p+1) \times 1$ vector of the parameters $(\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$.
- $\mathbf{E}$ is an $n \times 1$ vector of the error terms i.e, $\mathrm{E} = (\epsilon_1, \epsilon_2, \epsilon_3, \ldots, \epsilon_n)^T$.

5.3 Assumptions of Linear Regression

The Ordinary Least Squares (OLS) method, commonly used to estimate the parameters in a linear regression model, relies on several key assumptions:

1. **Linearity:** The relationship between independent and dependent variable is linear.

2. **Independence of Errors:** The error terms (ϵ_i) are independent of each other i.e that the error for one observation is not related to the error for any other observation.

3. **Homoscedasticity:** The error terms have constant variance across all the independent variables i.e the spread of the residuals is roughly the same for all predicted values.

4. **Mean varience of Errors:** The error terms have a mean of zero, $\mathbb{E}[\mathrm{E}] = \mathbf{0}$, and a constant variance, $\mathbb{D}(\mathrm{E}) = \sigma^2 I_n$

5. **Multicollinearity:** There is no perfect linear relationship among the independent variables.

6. **Normality of Errors:** The error terms are normally distributed. This assumption is particularly important for hypothesis testing and constructing confidence intervals.

5.4 Estimation of the β's

The most common method for estimating the parameters ($\beta_0, \beta_1, \ldots, \beta_p$) in a linear regression model is the Ordinary Least Squares (OLS) method. OLS chooses the values of the β's that minimize the sum of the squared differences between the observed and predicted values of the dependent variable.

The OLS estimator for β is given by:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{Y} \tag{3}$$

Where:

- $\hat{\beta}$ is the vector of estimated coefficients.
- $\mathbf{X}^T$ is the transpose of the matrix $\mathbf{X}$.
- $(\mathbf{X}^T\mathbf{X})^{-}$ is the moore-penrose g-inverse of the matrix $(\mathbf{X}^T\mathbf{X})$.

6 Model Fitting

After explanatory data analysis (EDA) and data processing, we aim to model each of the banks individually. Owing to the interpretable capability of the linear model, we will look to compare the regression result for each of the six banks.

6.1 Citigroup Modeling

In order to model the returns of Citigroup, we need to keep few things in mind. Here we will be using returns i.e. the ratio of today's stock price to yesterday's price. We are avoiding log returns as the banking sector is comparatively a stable industry so taking log returns of an already stable measure will diminish the value to a miniscule amount.

Again for modeling we are not using the values of the current macro-economic factors, i.e., for i^{th} time point we will not be using the macro-economic indicators of the same time point, because during the time of forecasting the current macro-economic values may not be available, we might only have to use the previous observations of the same. For consistency, we will be using upto the 4^{th} lag of all the bank data and also the macro-economic indicators. Hence,

$Y : Returns\ of\ Citigroup,$

$X : all\ 4\ lags\ of\ returns\ of\ the\ six\ banks\ along\ with\ the\ differenced\ values\ of\ the\ macroeconomic\ indicators.$

6.1.1 Full Model

In the first step, we look to fit a full model on the whole dataset. The model is:

$$Y = X\beta + \mathbf{E} \quad , where\ \beta\ is\ the\ coefficient\ vector\ and\ \mathbf{E} \sim\ N(0_n, \sigma^2 I_n)$$

The model is fitted using Least Squares Method and below we provide the model findings.

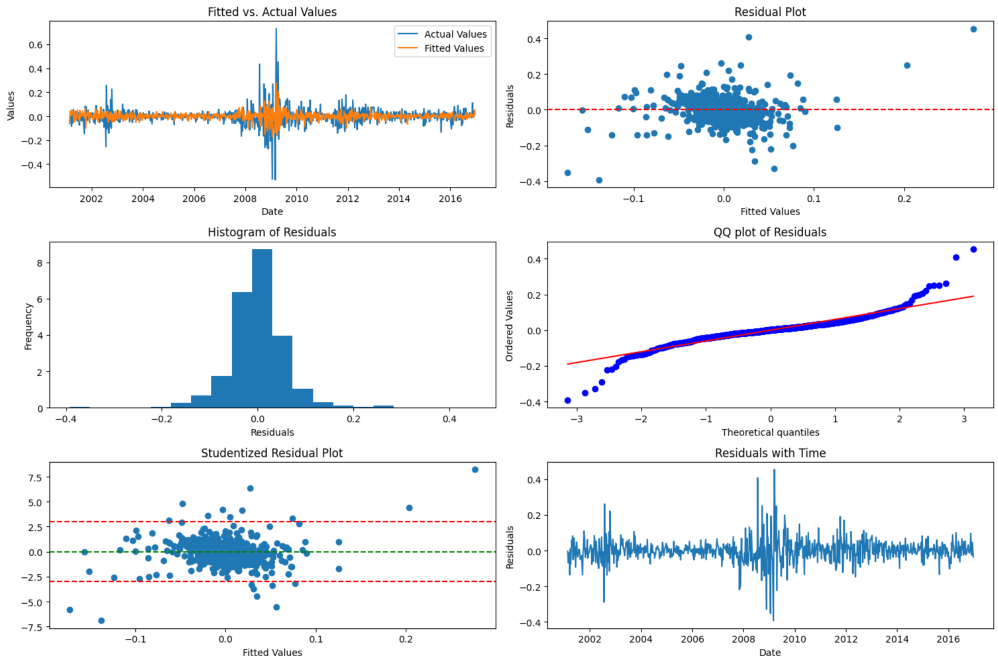
	$\mathbf{R^2}$	**Adj. $\mathbf{R^2}$**	**AIC**	**BIC**
Full Model	0.20	0.16	-2081	-1888

Full Model Performance Metrics

6.1.2 Backward Selection Model

Now we use the backward elimination method to get an optimized set of features for our regression. These are the findings from backward selection.

We have selected the following variables after backward selection:



Diagnostic Plots for Full Model

Variable Name		
Citigroup lag1	JPMorgan Chase lag1	Goldman Sachs lag1
Morgan Stanley lag1	Citigroup lag2	Bank of America lag2
JPMorgan Chase lag2	Morgan Stanley lag2	JPMorgan Chase lag3
Wells Fargo lag3	Morgan Stanley lag3	S&P500 diff lag3
Inflation diff lag3	Citigroup lag4	Bank of America lag4
JPMorgan Chase lag4	Wells Fargo lag4	Morgan Stanley lag4
Federal Funds Rate diff lag4		

Table 1: List of Variables from Regression Summary

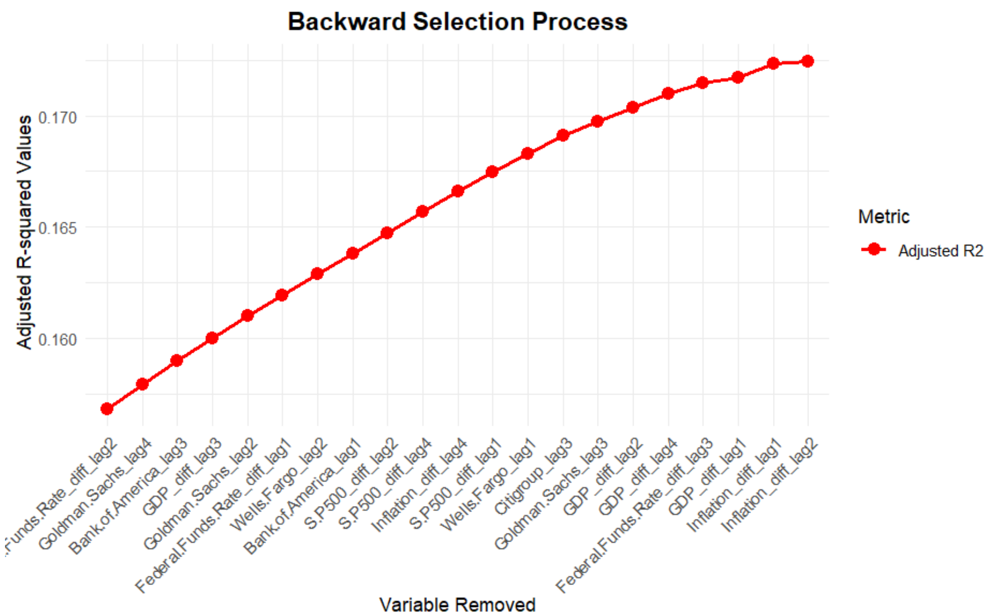
The model is fitted using Least Squares Method and below we provide the model findings.

We have eliminated 21 variables and just lost 0.01 in adjusted R^2, so backward selection gives a relevant choice of predictors without considerable decrease in the efficiency of the model.

6.1.3 Forward Selection Model

Now we use the forward elimination method to get an optimized set of features for our regression. These are the findings from forward selection.

We have selected the following variables after backward selection:



	R^2	Adj. R^2	AIC	BIC
Full Model	0.20	0.16	-2081	-1888
Backward Elimination	0.19	0.17	-2118	-2024

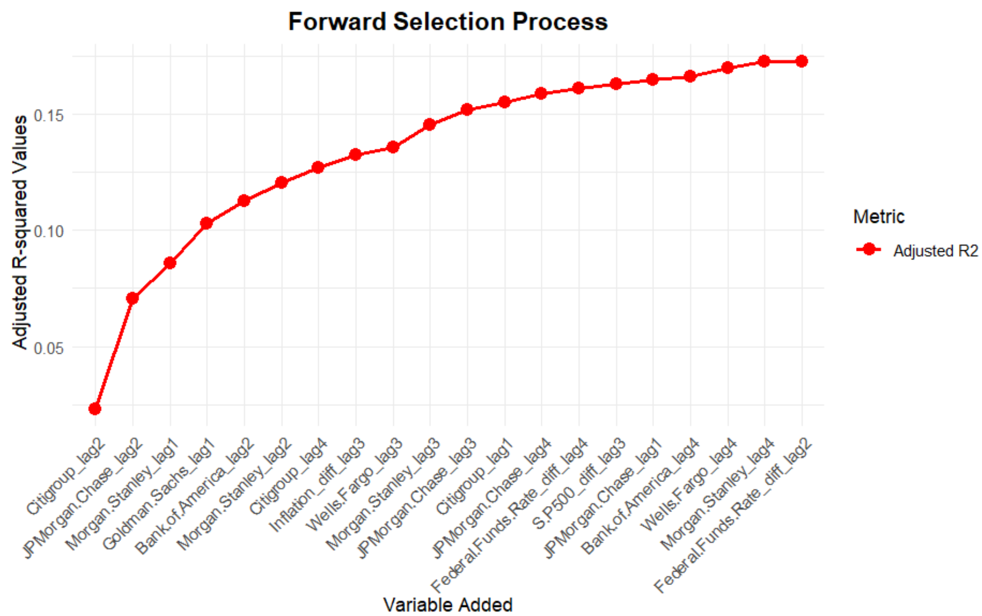
Comparing Model Performance Metrics

Variable Name		
Citigroup lag2	JPMorgan Chase lag2	Morgan Stanley lag1
Goldman Sachs lag1	Bank of America lag2	Morgan Stanley lag2
Citigroup lag4	Inflation diff lag3	Wells Fargo lag3
Morgan Stanley lag3	JPMorgan Chase lag3	Citigroup lag1
JPMorgan Chase lag4	Federal Funds Rate diff lag4	SP500 diff lag3
JPMorgan Chase lag1	Bank of America lag4	Wells Fargo lag4
Morgan Stanley lag4	Federal Funds Rate diff lag2	

Table 2: List of Variables from Regression Summary

The model is fitted using Least Squares Method and below we provide the model findings.

From the forward and backward selection method we see that backward elimination achieves the most reduction in feature space. So we use the features selected after backward elimination moving forward.



	R^2	Adj. R^2	AIC	BIC
Full Model	0.20	0.16	-2081	-1888
Backward Elimination	0.19	0.17	-2118	-2024
Forward Selection	0.18	0.16	-2101	-2001

Comparing Model Performance Metrics

6.1.4 Model Diagnostics of Backward Eliminated Model

In this section, we take a deeper look at the backward eliminated model by performing the model diagnostics and testing for the assumptions of the model. This step is important because the linear model and the estimates of the coefficients will have many desirable properties if it abides by the assumptions. First we give the higher level overview of our model fit and the residuals we obtained from the model.

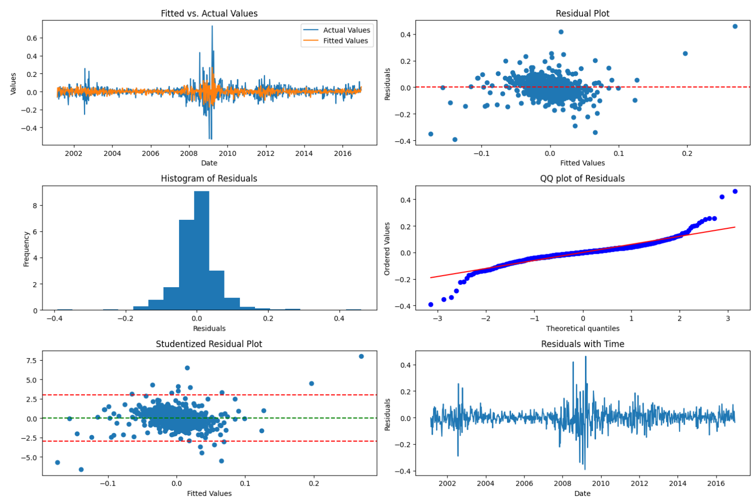
From a cursory glance we can observe that our model has the capability to predict the movements of the market for a moderate scale, for movements with high magnitudes it is not fully capture the full market dynamics. The residuals also seem to be normally distributed and the residuals seem to be clustered around the middle with some notable high values.

In order to give statistically backed conclusions we need to perform statistical tests of these possible observations. This takes us to testing our assumptions.

We use the Kolmogorov-Smirnov Test for testing for normality test of the residuals, the Llunj-Box test for autocorrelation, checked the Variance Inflation Factor (VIF) for

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0004	0.002	-0.187	0.851	-0.005	0.004
Citigroup_lag1	0.1340	0.050	2.687	0.007	0.036	0.232
JPMorgan Chase_lag1	-0.1563	0.079	-1.968	0.049	-0.312	-0.000
Goldman Sachs_lag1	0.4771	0.098	4.851	0.000	0.284	0.670
Morgan Stanley_lag1	-0.4612	0.075	-6.171	0.000	-0.608	-0.315
Citigroup_lag2	-0.5025	0.061	-8.186	0.000	-0.623	-0.382
Bank of America_lag2	0.3099	0.068	4.572	0.000	0.177	0.443
JPMorgan Chase_lag2	0.4171	0.080	5.244	0.000	0.261	0.573
Morgan Stanley_lag2	-0.1649	0.057	-2.871	0.004	-0.278	-0.052
JPMorgan Chase_lag3	-0.2360	0.082	-2.873	0.004	-0.397	-0.075
Wells Fargo_lag3	-0.0956	0.079	-1.212	0.226	-0.250	0.059
Morgan Stanley_lag3	0.2767	0.056	4.902	0.000	0.166	0.387
S&P500_diff_lag3	-0.0002	7.88e-05	-2.062	0.040	-0.000	-7.78e-06
Inflation_diff_lag3	0.0998	0.040	2.481	0.013	0.021	0.179
Citigroup_lag4	0.1547	0.063	2.447	0.015	0.031	0.279
Bank of America_lag4	0.1933	0.074	2.616	0.009	0.048	0.338
JPMorgan Chase_lag4	-0.1255	0.088	-1.434	0.152	-0.297	0.046
Wells Fargo_lag4	-0.2127	0.092	-2.300	0.022	-0.394	-0.031
Morgan Stanley_lag4	-0.1076	0.058	-1.855	0.064	-0.221	0.006
Federal Funds Rate_diff_lag4	0.0156	0.008	1.853	0.064	-0.001	0.032

Model Summary of Backward Selection Model



Diagnostic plots of the backward elimination model

multicollinearity.

Assumption	**Test**	**Remark**
Multicollinearity	VIF	There is no multicollinearity
Normality	Kolmogorov-Smirnov Test	The residuals are non-normal
Homoscedasticity	Breausch-Pagan Test	The test fails hinting existence of heteroscedasicity
Auto-correlation	Llunj Box Text	The test shows no auto correlation

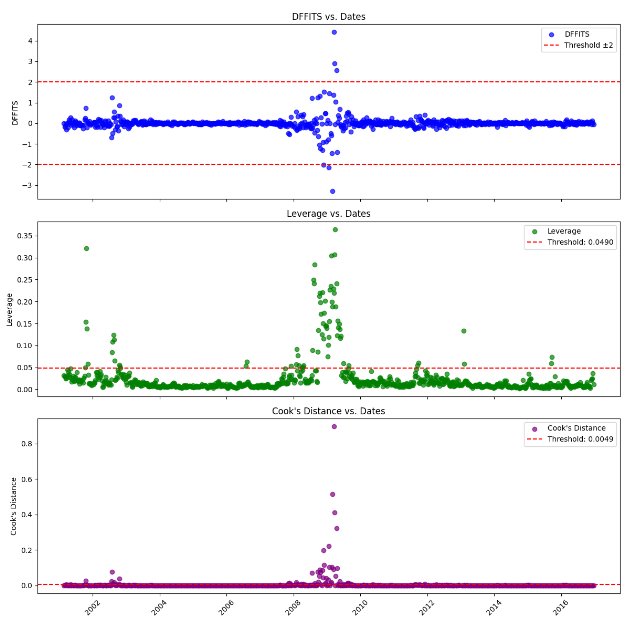
Model Assumptions

Here we see that there are two assumptions that are failing in our model, firstly, the non-normal residuals and secondly the heteroscedasticity. To battle the non-normal residual problem, we look to transform the already transformed y variable, we apply box cox transformation and get the best lambda to be close to 1, which implies no such transformation at all. To mend the heteroscedasticity, we can use generalised least squares or a simpler version of it, the weighted least squares. Another route that can be taken is a robust estimate for the standard errors, i.e., Heteroskedasticity-consistent (HC) standard errors.

Here we provide the HC standard errors estimate for our model.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0004	0.003	-0.166	0.868	-0.006	0.005
Citigroup_lag1	0.1340	0.157	0.852	0.394	-0.175	0.443
JPMorgan Chase_lag1	-0.1563	0.202	-0.774	0.439	-0.553	0.240
Goldman Sachs_lag1	0.4771	0.179	2.673	0.008	0.127	0.828
Morgan Stanley_lag1	-0.4612	0.161	-2.857	0.004	-0.778	-0.144
Citigroup_lag2	-0.5025	0.193	-2.608	0.009	-0.881	-0.124
Bank of America_lag2	0.3099	0.108	2.880	0.004	0.099	0.521
JPMorgan Chase_lag2	0.4171	0.227	1.836	0.067	-0.029	0.863
Morgan Stanley_lag2	-0.1649	0.118	-1.394	0.164	-0.397	0.067
JPMorgan Chase_lag3	-0.2360	0.148	-1.600	0.110	-0.526	0.054
Wells Fargo_lag3	-0.0956	0.149	-0.641	0.522	-0.388	0.197
Morgan Stanley_lag3	0.2767	0.110	2.515	0.012	0.061	0.493
S&P500_diff_lag3	-0.0002	9.6e-05	-1.692	0.091	-0.000	2.6e-05
Inflation_diff_lag3	0.0998	0.056	1.771	0.077	-0.011	0.210
Citigroup_lag4	0.1547	0.151	1.022	0.307	-0.143	0.452
Bank of America_lag4	0.1933	0.186	1.041	0.298	-0.171	0.558
JPMorgan Chase_lag4	-0.1255	0.135	-0.930	0.352	-0.390	0.139
Wells Fargo_lag4	-0.2127	0.261	-0.815	0.415	-0.725	0.299
Morgan Stanley_lag4	-0.1076	0.118	-0.911	0.363	-0.340	0.124
Federal Funds Rate_diff_lag4	0.0156	0.011	1.362	0.174	-0.007	0.038

Model Summary of Backward Selection Model



Leverage point Analysis

Next we look to see the influential points of this model. We use Cooks Distance, leverages, DFFITS and DFBETAS. We provide the plots for the same except the later due to space reasons.

From the leverage point plots we see that the finanial crisis episode has produced the most amounts of outliers. In our next steps we look to battle with this challenge in our modeling process. We have come up with a few ways to solve this problem, some of these have improved our results significantly, while some has not. In the next few sections we try to introduce these ideas.

In the next section we try to exclude the part with the highest volatility in the years 2008 and 2009 and try to fit two different linear regression models on the two parts and see how that works.

6.1.5 Weighted Least Squares

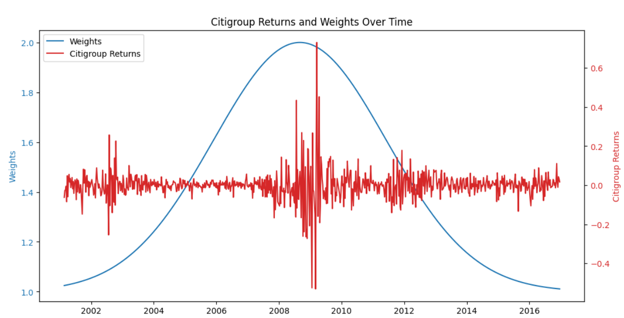
We split the data into two parts avoiding a the highly volatile part, and subsequently modeling the two timelines as independent models. In this section we look to use weighted least squares to model the varying nature of the variance of the errors.

We put forward a few observations, before introducing a variance structure for our model. At times of high volatility, we see that the market starts ti experience rapid currents before eventually tanking, this means the actual drop is not just a one-day phenomenon, rather it just builds up under the rug after eventually the disaster happens, we are looking to mimic that nature of the market at 2008 but taking a positive function whose values increase as it approaches the 2008 financial crisis and then reaches its pinnacle at the day of mayhem and then starts to decline gradually. We use a guassian type weighting scheme to implement it.

$weight_t = exp\{-\frac{l_t^2}{2\sigma^2}\}$, where l_t is the difference in no. of days from Sept. 2008

$$and\ \sigma^2 = max\{first\ date\ of\ the\ day - Sept.2008,\ Sept.2008 - \ last\ date\ of\ the\ data\}/3$$

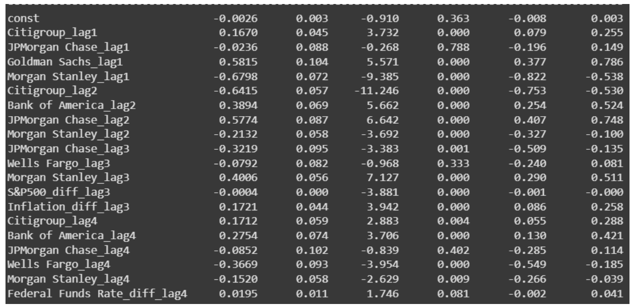
Also we added a 1 to all the weights to avoid having 0 weights. There is no particular reason of using a symmetric function to model the variance structure of the market, we are using it as a baseline. In reality the actual dynamics of the model is more asymmetric or even more complex. We provide the dual axis chart for the weights and the Citigroup returns.



Varying Weights

Based on these weights we use the weighted least squares method to fit the linear regression model with Citigroup model as our y and the variables selected in backward elimination method as our predictors.

Now we present the model findings from this weighted model.



const	-0.0026	0.003	-0.910	0.363	-0.008	0.003
Citigroup_lag1	0.1670	0.045	3.732	0.000	0.079	0.255
JPMorgan Chase_lag1	-0.0236	0.088	-0.268	0.788	-0.196	0.149
Goldman Sachs_lag1	0.5815	0.104	5.571	0.000	0.377	0.786
Morgan Stanley_lag1	-0.6798	0.072	-9.385	0.000	-0.822	-0.538
Citigroup_lag2	-0.6415	0.057	-11.246	0.000	-0.753	-0.530
Bank of America_lag2	0.3894	0.069	5.662	0.000	0.254	0.524
JPMorgan Chase_lag2	0.5774	0.087	6.642	0.000	0.407	0.748
Morgan Stanley_lag2	-0.2132	0.058	-3.692	0.000	-0.327	-0.100
JPMorgan Chase_lag3	-0.3219	0.095	-3.383	0.001	-0.509	-0.135
Wells Fargo_lag3	-0.0792	0.082	-0.968	0.333	-0.240	0.081
Morgan Stanley_lag3	0.4006	0.056	7.127	0.000	0.290	0.511
S&P500_diff_lag3	-0.0004	0.000	-3.881	0.000	-0.001	-0.000
Inflation_diff_lag3	0.1721	0.044	3.942	0.000	0.086	0.258
Citigroup_lag4	0.1712	0.059	2.883	0.004	0.055	0.288
Bank of America_lag4	0.2754	0.074	3.706	0.000	0.130	0.421
JPMorgan Chase_lag4	-0.0852	0.102	-0.839	0.402	-0.285	0.114
Wells Fargo_lag4	-0.3669	0.093	-3.954	0.000	-0.549	-0.185
Morgan Stanley_lag4	-0.1520	0.058	-2.629	0.009	-0.266	-0.039
Federal Funds Rate_diff_lag4	0.0195	0.011	1.746	0.081	-0.002	0.041

Varying Weights

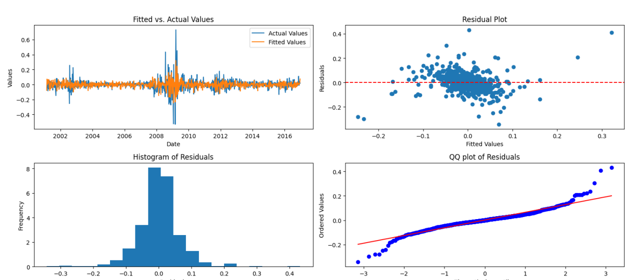
We summarize the important metrics of this model in this table:

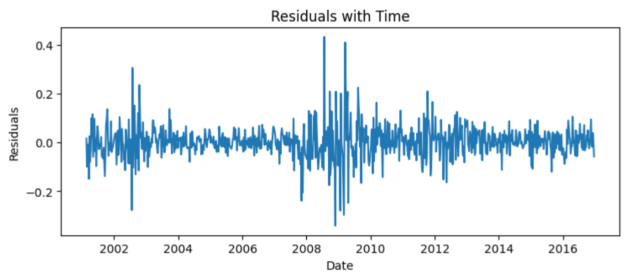
	R^2	Adj. R^2	AIC	BIC
Full Model	0.20	0.16	-2081	-1888
Backward Elimination	0.19	0.17	-2118	-2024
Forward Selection	0.18	0.16	-2101	-2001
Weighted Least Squares(WLS)	0.319	0.303	-1343	-1249

Comparing Model Performance Metrics

This model performs quite better than the multiple linear regression model.

We also give the diagnostic plots for the same. From the plot we observe that this model tries to capture the high volatility more in the 2008 timeline, this was exactly the idea behind choosing our weights in that manner.





Diagnostic Plots

We also test for the assumptions

Assumption	**Test**	**Remark**
Multicollinearity	VIF	There is no multicollinearity
Normality	Kolmogorov-Smirnov Test	The residuals are non-normal
Homoscedasticity	Breausch-Pagan Test	The test fails hinting existence of heteroscedasicity
Auto-correlation	Llunj Box Text	The test shows no auto correlation

Model Assumptions

Even after taking a different model for the variances we still have heteroskedasticity, which we are not surprised by as the underlying variance model of the stock returns is much more complex. Now we look to take multiple crucial dates in the weights structure, making the variance model much more robust. Here we present the few financially important dates that can be important to explain some of the variable nature.

So we use a kind of mixture of effect functions and fit a model using them as the weights, also one thing to consider, at a time point t the immediate future event has the only effect, depending on its distance from the event date. As the events after that is very unlikely to have been brewing underneath before that.

Now we provide the weights with the Citigroup returns

Based on these weights we use the weighted least squares method to fit the linear regression model with Citigroup model as our y and the variables selected in backward elimination method as our predictors.

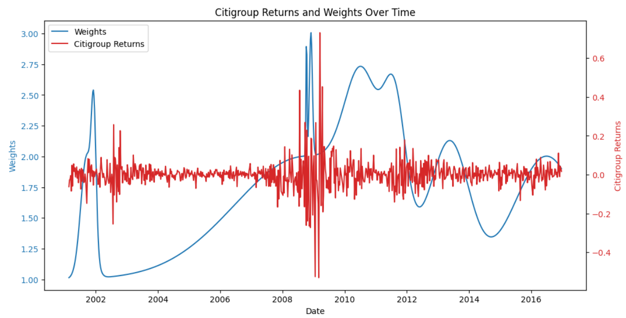
Now we present the model findings from this weighted model.

We summarize the important metrics of this model in this table:

Well this model didn't perform well as anticipated. This also gives us a reality check that all intuitive models do not always work. It performed better than the backward selection model but the improved performance is not worth the added complexity.

Event	Date	Description
9/11 Attacks	2001-09-11	Caused severe market panic and uncertainty, leading to a sharp decline in stock prices due to fears of economic slowdown and losses in insurance and aviation-related sectors.
Enron Bankruptcy	2001-12-02	Shook investor confidence in corporate governance and accounting transparency, affecting the credibility of financial reporting and impacting banks exposed to similar risks.
Lehman Collapse	2008-09-15	Triggered a systemic financial crisis with liquidity shortages and interbank distrust, severely impacting global banking stocks.
TARP Bailout	2008-10-03	Provided capital support to major banks, stabilizing short-term investor sentiment but raising concerns about long-term moral hazard and regulation.
Fed QE1 Announcement	2008-11-25	Signaled aggressive monetary easing to support credit markets, boosting liquidity and improving bank stock valuations.
Dodd-Frank Act Signed	2010-07-21	Imposed stricter financial regulations on banks, potentially reducing profitability but enhancing long-term system stability.
US Debt Downgrade	2011-08-05	Led to broad market volatility and concerns over U.S. fiscal health, negatively affecting risk sentiment across bank stocks.
Taper Tantrum	2013-05-22	Rising interest rate expectations due to Fed tapering talks increased funding cost concerns and market volatility, affecting bank valuations.
Brexit Vote	2016-06-23	Introduced geopolitical and economic uncertainty, especially impacting globally exposed and UK-linked banks due to fears of regulatory divergence.

Major Events and Their Impact on Bank Stock Returns



Varying Weights for multiple models

const	-0.0008	0.002	-0.323	0.747	-0.006	0.004
Citigroup_lag1	0.1473	0.048	3.050	0.002	0.053	0.242
JPMorgan Chase_lag1	-0.1592	0.083	-1.906	0.057	-0.323	0.005
Goldman Sachs_lag1	0.5605	0.099	5.677	0.000	0.367	0.754
Morgan Stanley_lag1	-0.5114	0.074	-6.952	0.000	-0.656	-0.367
Citigroup_lag2	-0.5447	0.059	-9.162	0.000	-0.661	-0.428
Bank of America_lag2	0.3300	0.068	4.843	0.000	0.196	0.464
JPMorgan Chase_lag2	0.4611	0.083	5.543	0.000	0.298	0.624
Morgan Stanley_lag2	-0.1776	0.057	-3.138	0.002	-0.289	-0.067
JPMorgan Chase_lag3	-0.2711	0.088	-3.093	0.002	-0.443	-0.099
Wells Fargo_lag3	-0.0936	0.080	-1.174	0.241	-0.250	0.063
Morgan Stanley_lag3	0.3044	0.055	5.534	0.000	0.196	0.412
S&P500_diff_lag3	-0.0002	8.13e-05	-2.524	0.012	-0.000	-4.56e-05
Inflation_diff_lag3	0.1165	0.041	2.824	0.005	0.036	0.197
Citigroup_lag4	0.1417	0.062	2.293	0.022	0.020	0.263
Bank of America_lag4	0.2148	0.073	2.926	0.004	0.071	0.359
JPMorgan Chase_lag4	-0.0798	0.094	-0.849	0.396	-0.264	0.105
Wells Fargo_lag4	-0.2656	0.091	-2.912	0.004	-0.445	-0.087
Morgan Stanley_lag4	-0.1226	0.057	-2.134	0.033	-0.235	-0.010
Federal Funds Rate_diff_lag4	0.0151	0.008	1.837	0.067	-0.001	0.031

Varying Weights

If we used some asymmetric function, we might be able to mimic the market movements better.

6.1.6 Using Regression on Two timelines

Now based on the analysis performed in the previous section, we will use the regression on two different parts of the dataset to see how well our model is performing. Hence, we divide the data on the crisis that occurred in September 2008, widely known as the collapse of *Lehman Brothers.*

We will begin the modeling of the first part of the data i.e. the part before the crisis happened. As in the previous section, we have done the forward and backward selection on the trimmed data and get the backward selection method performs well. At first we will see the overall overview of the data, and then we will see the diagnostics of the data. model.

Here are the model summary of the Backward Elimination Model as performed,

Now we will check the different assumptions of the model such as normality, multicollinearity, etc. For such checking we will use some statistical tests like the Kolmogorov-Smirnov test for testing for normality test of the residuals, the Llunj-Box test for autocorrelation, and checked the Variance Inflation Factor (VIF) for multicollinearity.

	$\mathbf{R^2}$	**Adj. $\mathbf{R^2}$**	**AIC**	**BIC**
Full Model	0.20	0.16	-2081	-1888
Backward Elimination	0.19	0.17	-2118	-2024
Forward Selection	0.18	0.16	-2101	-2001
Weighted Least Squares(WLS)	0.319	0.303	-1343	-1249
WLS with multiple events	0.22	0.20	-2005	-1911

Comparing Model Performance Metrics

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0004	0.002	0.160	0.873	-0.004	0.005
Citigroup_lag1	-0.1909	0.080	-2.377	0.018	-0.349	-0.033
Wells Fargo_lag1	-0.3270	0.095	-3.440	0.001	-0.514	-0.140
S&P500_diff_lag1	0.0002	9.51e-05	2.533	0.012	5.39e-05	0.000
GDP_diff_lag1	-0.0030	0.001	-2.162	0.031	-0.006	-0.000
Inflation_diff_lag1	0.1788	0.057	3.159	0.002	0.067	0.290
Federal Funds Rate_diff_lag1	-0.0290	0.011	-2.739	0.006	-0.050	-0.008
Citigroup_lag2	0.2224	0.092	2.425	0.016	0.042	0.403
Bank of America_lag2	0.1627	0.102	1.599	0.111	-0.037	0.363
Wells Fargo_lag2	-0.3220	0.123	-2.614	0.009	-0.564	-0.080
GDP_diff_lag2	0.0021	0.001	1.423	0.155	-0.001	0.005
Inflation_diff_lag2	-0.1355	0.064	-2.124	0.034	-0.261	-0.010
Federal Funds Rate_diff_lag2	-0.0357	0.011	-3.242	0.001	-0.057	-0.014
Bank of America_lag3	-0.1147	0.059	-1.939	0.053	-0.231	0.002
Goldman Sachs_lag3	-0.2104	0.103	-2.046	0.041	-0.412	-0.008
Morgan Stanley_lag3	0.1612	0.083	1.943	0.053	-0.002	0.324
GDP_diff_lag3	0.0032	0.001	2.278	0.023	0.000	0.006
Inflation_diff_lag3	-0.0879	0.057	-1.544	0.123	-0.200	0.024
Bank of America_lag4	-0.3323	0.092	-3.627	0.000	-0.512	-0.152
JPMorgan Chase_lag4	-0.2675	0.089	-3.012	0.003	-0.442	-0.093
Wells Fargo_lag4	0.3761	0.120	3.124	0.002	0.139	0.613
Morgan Stanley_lag4	0.1739	0.072	2.408	0.017	0.032	0.316
GDP_diff_lag4	0.0016	0.001	1.337	0.182	-0.001	0.004
Federal Funds Rate_diff_lag4	0.0175	0.007	2.395	0.017	0.003	0.032

Model summary of the Backward Model

Here we see that there are two assumptions that are failing in our model, firstly, the non-normal residuals and secondly the heteroscedasticity. To battle the non-normal residual problem, we look to transform the already transformed y variable, we apply box cox transformation and get the best lambda to be close to 1, which implies no such transformation at all. To mend the heteroscedasticity, we can use generalised least squares or a simpler version of it, the weighted least squares. Another route that can be taken is a robust estimate for the standard errors, i.e., Heteroskedasticity-consistent (HC) standard errors.Hence we provide the HC standard errors estimate for our model.

Also here are some diagnostics plot of the model,

Next we look to see the influential points of this model. We use Cooks Distance, leverages, DFFITS. We provide the plots for the same.

Here we can see that, from the plots, the deviations are just at the starting points and the plot is improved from the previous model.

Now we will start doing the modelling for the later part i.e after the crisis. We will do the

	$\mathbf{R^2}$	**Adj.** $\mathbf{R^2}$	**AIC**	**BIC**
Backward Elimination	0.249	0.201	-1260	-1165

Assumption	**Test**	**Remark**
Multicollinearity	VIF	There is no multicollinearity
Normality	Kolmogorov-Smirnov Test	The residuals are non-normal
Homoscedasticity	Breausch-Pagan Test	Heteroscedasicity is present
Auto-correlation	Llunj Box Text	The test shows no auto correlation

Model Assumptions

same procedure as done in the previous section . This time also the Backward Elimination model performs better than the forward selection model. At first we will model on the selected variables i.e

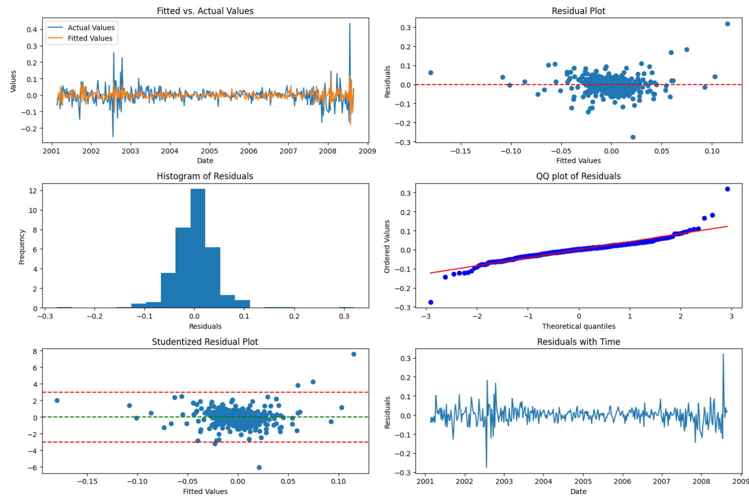
Variable Name		
JPMorgan Chase lag1	Inflation diff lag2	Wells.Fargo lag3
Morgan.Stanley lag3	JPMorgan.Chase lag3	Bank.of.America ag4
Morgan.Stanley lag4	Goldman.Sachs lag2	S.P500 diff lag1
Wells.Fargo lag1	Wells.Fargo lag4	Morgan.Stanley lag1
Goldman.Sachs lag1	Bank.of.America lag3	Goldman.Sachs lag4
S.P500 diff lag2	Morgan.Stanley lag2	S.P500 diff lag4

List of Variables from Selected for backward Elimination

Now, here is the model summary and the coefficients; later, we will try to validate the model assumptions.

Now we will check the different assumptions of the model such as normality, multicollinearity, etc as done previously. For such checking, we will use some statistical tests like the Kolmogorov-Smirnov test for testing for normality test of the residuals, the Llunj-Box test for autocorrelation, and checked the Variance Inflation Factor (VIF) for multicollinearity.

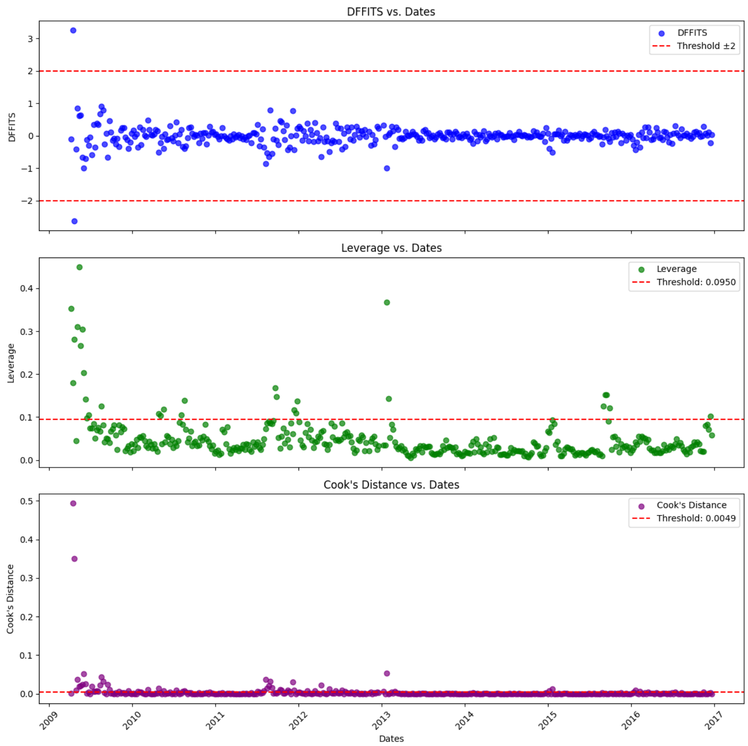
Here we see that there are two assumptions that are failing in our model, firstly, the non-normal residuals and secondly the heteroscedasticity. To battle the non-normal residual problem, we look to transform the already transformed y variable, we apply box cox transformation and get the best lambda to be close to 1, which implies no such transformation at all. To mend the heteroscedasticity, we can use generalised least squares or a simpler version of it, the weighted least squares. Another route that can be taken is a robust estimate for the standard errors, i.e., Heteroskedasticity-consistent (HC) standard errors. Here we provide the HC standard errors estimate for our model.



Diagnostic plots of the backward elimination model

$\mathbf{R^2}$	**Adj. $\mathbf{R^2}$**	**AIC**	**BIC**
0.126	0.085	-1191	-1115

Also here are the diagnostic plots of the model



Leverage point Analysis

Assumption	**Test**	**Remark**
Multicollinearity	VIF	There is no multicollinearity
Normality	Kolmogorov-Smirnov Test	The residuals are non-normal
Homoscedasticity	Breausch-Pagan Test	Heteroscedasicity is present
Auto-correlation	Llunj Box Text	The test shows no auto correlation

Model Assumptions

Next we look to see the influential points of this model. We use Cooks Distance, leverages, DFFITS. We provide the plots for the same.

Here we can see that, from the plots, the deviations are a little bit scattered than the pre 2008 data. Hence we can give a small conclusion that the data is a bit less predictable for the later half.

Now we can draw some insights that are based on the pre and post crisis model.

- Pre-Crisis model is explaining more varience than the post-crisis model.
- Indicates structural and behavioral changes in financial markets after the 2008 crisis.

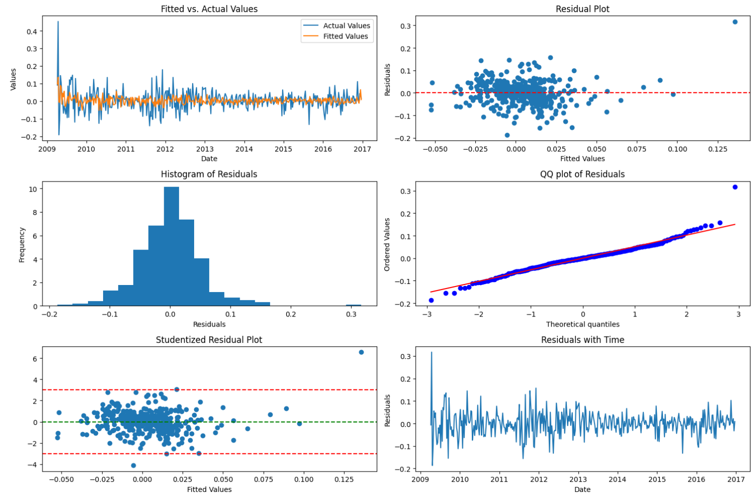
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0019	0.003	0.683	0.495	-0.004	0.007
JPMorgan Chase_lag1	-0.0310	0.154	-0.202	0.840	-0.334	0.271
Inflation_diff_lag2	0.0862	0.053	1.612	0.108	-0.019	0.191
Wells Fargo_lag3	0.0597	0.129	0.461	0.645	-0.195	0.314
Morgan Stanley_lag3	-0.0984	0.092	-1.064	0.288	-0.280	0.083
JPMorgan Chase_lag3	0.3124	0.142	2.195	0.029	0.033	0.592
Bank of America_lag4	0.0612	0.079	0.771	0.441	-0.095	0.217
Morgan Stanley_lag4	-0.2980	0.104	-2.873	0.004	-0.502	-0.094
Goldman Sachs_lag2	0.0110	0.134	0.082	0.935	-0.252	0.274
S&P500_diff_lag1	1.265e-05	0.000	0.126	0.900	-0.000	0.000
Wells Fargo_lag1	-0.2496	0.120	-2.085	0.038	-0.485	-0.014
Wells Fargo_lag4	0.2867	0.117	2.455	0.015	0.057	0.516
Morgan Stanley_lag1	-0.1905	0.108	-1.771	0.077	-0.402	0.021
Goldman Sachs_lag1	0.3799	0.141	2.689	0.007	0.102	0.658
Bank of America_lag3	-0.0571	0.082	-0.693	0.489	-0.219	0.105
Goldman Sachs_lag4	0.3566	0.130	2.745	0.006	0.101	0.612
S&P500_diff_lag2	0.0001	0.000	1.439	0.151	-5.33e-05	0.000
Morgan Stanley_lag2	-0.1647	0.099	-1.656	0.098	-0.360	0.031
S&P500_diff_lag4	-6.572e-05	6.97e-05	-0.943	0.346	-0.000	7.13e-05

Model summary of the Backward Model

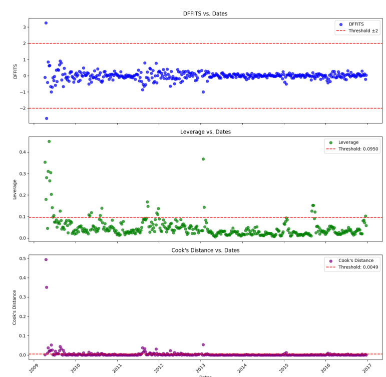
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0019	0.003	0.696	0.487	-0.003	0.007
JPMorgan Chase_lag1	-0.0310	0.174	-0.178	0.859	-0.373	0.311
Inflation_diff_lag2	0.0862	0.067	1.290	0.198	-0.045	0.218
Wells Fargo_lag3	0.0597	0.170	0.351	0.726	-0.275	0.394
Morgan Stanley_lag3	-0.0984	0.109	-0.903	0.367	-0.313	0.116
JPMorgan Chase_lag3	0.3124	0.149	2.091	0.037	0.019	0.606
Bank of America_lag4	0.0612	0.101	0.606	0.545	-0.137	0.260
Morgan Stanley_lag4	-0.2980	0.173	-1.721	0.086	-0.638	0.042
Goldman Sachs_lag2	0.0110	0.155	0.071	0.944	-0.294	0.316
S&P500_diff_lag1	1.265e-05	0.000	0.120	0.905	-0.000	0.000
Wells Fargo_lag1	-0.2496	0.195	-1.278	0.202	-0.633	0.134
Wells Fargo_lag4	0.2867	0.212	1.355	0.176	-0.129	0.703
Morgan Stanley_lag1	-0.1905	0.149	-1.281	0.201	-0.483	0.102
Goldman Sachs_lag1	0.3799	0.239	1.593	0.112	-0.089	0.849
Bank of America_lag3	-0.0571	0.119	-0.478	0.633	-0.292	0.178
Goldman Sachs_lag4	0.3566	0.188	1.902	0.058	-0.012	0.725
S&P500_diff_lag2	0.0001	0.000	1.317	0.189	-7.18e-05	0.000
Morgan Stanley_lag2	-0.1647	0.129	-1.272	0.204	-0.419	0.090
S&P500_diff_lag4	-6.572e-05	8.1e-05	-0.811	0.418	-0.000	9.36e-05

Diagnostic plots of the backward elimination model

- The pre-crisis model shows stronger and more intuitive relationships between macroeconomic indicators, peer bank performance, and Citigroup's returns.



Diagnostic plots of the backward elimination model



Leverage point Analysis

Conclusion

In this study, we analyzed the return behavior of major American banks using various macroeconomic indicators and inter-bank index lags.Among the models tested, the Weighted Least Squares (WLS) model outperformed others with the highest R^2 and lowest information criteria values. Our findings highlight that incorporating economic volatility through weighted structures yields more accurate and robust predictions.Lag values of the bank indices also proved to be significant, underscoring the interconnected nature of the financial sector.These insights contribute to a better understanding of systemic risk and bank return predictability during turbulent periods. Overall, the project supports the use of dynamic modeling techniques for informed financial decision-making.

7 Scope for Further Research

While our project covers significant ground, several promising extensions were left unexplored due to **constraints** of **time** and **technical expertise**. We plan to revisit and enhance the study in the following ways:

1. **Inclusion of Additional Predictors:** The current model includes six leading banks and four macroeconomic indicators. Incorporating more financial institutions and variables could improve model performance and coverage.

2. **Modeling Each Bank Separately:** Thus far, only Citigroup has been fully modeled, with partial progress on JPMorgan Chase. Developing distinct models for the remaining banks—Bank of America, Wells Fargo, Goldman Sachs, and Morgan Stanley—will offer institution-specific insights.

3. **Optimizing Weighted Least Squares (WLS):** The weights used in our current WLS model are basic. Fine-tuning them based on residual variance or volatility patterns could lead to higher R^2 values and more accurate estimation.

4. **Event Date Hyperparameter Selection:** We intend to treat crisis dates (from Chow tests and real events) as hyperparameters. Evaluating different combinations will help identify the most influential time points for improved model accuracy.

5. **Handling Outliers via Quantile Regression:** Our dataset contains numerous outliers, which are crucial and cannot be simply removed. Employing quantile regression in future work would allow us to better capture the impact of macroeconomic factors across different parts of the return distribution.

6. **Stress Testing Scenarios:** Incorporating formal stress testing techniques can help analyze how extreme changes in macroeconomic indicators may affect bank stock performance and systemic risk.

7. **Grouped Lasso for Institutional Types:** Using grouped lasso by categorizing banks into Commercial, Investment, and Insurance types could enable better feature selection and reveal structural differences among these groups.

These enhancements would refine the current approach and lead to a more comprehensive understanding of the financial sector's response to macroeconomic dynamics and crises.

8 Acknowledgement

We sincerely thank our project supervisor, **Dr. Monika Bhattacharjee**, for their constant guidance and support. We are also grateful to our department, peers, and families for their encouragement throughout this project.