

|  |
| --- |
| Regression Analysis of |
| Financial Stock Returns |

# Introduction

The application of statistical regression analysis in financial markets has been a cornerstone of investment strategies and financial decision-making. This project explores the dynamics of stock prices through regression models, aiming to uncover patterns and try to predict future movements with higher accuracy. The study focuses on understanding the movement of Amazon Inc.’s stock price based on several factors. A select group of technology and index stocks, including SPY (S&P 500 ETF), AAPL (Apple Inc.), MSFT (Microsoft Corporation), and GOOGL (Alphabet Inc.) are used as the primary regressors. These stocks were chosen due to their significant impact on the market and their representation of both sector-specific and market-wide trends.

The analysis also explores the use of risk related factors through Fama French factors to see if it helps to better understand the fluctuations in Amazon's stock price.

The primary objective of this analysis is to apply various regression techniques to understand how well different models capture the returns of these stocks relative to market movements and other factors. By investigating the predictive capabilities of these models, this research contributes to the broader academic and practical understanding of asset pricing and investment strategy optimization in the volatile environment of financial markets.

# About Dataset

Dataset has two parts.

1. Stock Returns of Various Companies.

Stock prices in the code are sourced from Yahoo Finance, a renowned platform offering a wealth of financial information including stock quotes, financial news, press releases, and financial reports. The Python library yfinance is utilised to download this historical market data conveniently.

The specific type of stock price chosen is the Adjusted Close Price.

Adjusted Close Price: This represents the closing price of a stock after adjustments have been made for corporate actions such as dividends, stock splits, and new stock offerings. This price provides a more accurate picture of the stock’s value over time. It is the most relevant price when looking at historical returns or conducting a detailed analysis of past performance.

We are using the adjusted closing prices of the stocks ‘SPY’, ‘AAPL’, ‘AMZN’, ‘MSFT’, ‘GOOGL’ specified for each trading day from January 1, 2015, to December 31, 2023.

2. Fama-French Factor Data

The second part of the dataset is the Fama French Factors of the S&P 500 index, and for each day, we have the stock price. The Fama-French factors are a set of factors that explain the returns of a stock or a portfolio. Here the five factors are constructed using the 6 value-weighted portfolios formed on size and book-to-market, the 6 valueweight portfolios formed on size and operating profitability, and the 6 value-weight portfolios formed on size and investment.

* SMB: SMB measures the return on a diversified portfolio of small-cap stocks minus the return on a diversified portfolio of big-cap stocks.

SMB (B/M) = 1/3 (Small Value + Small Neutral + Small Growth) - 1/3 (Big

Value + Big Neutral + Big Growth)

SMB (OP) = 1/3 (Small Robust + Small Neutral + Small Weak) - 1/3 (Big

Robust + Big Neutral + Big Weak)

SMB (INV) = 1/3 (Small Conservative + Small Neutral + Small Aggressive)

- 1/3 (Big Conservative + Big Neutral + Big Aggressive)

SMB = 1/3 (SMB (B/M) + SMB (OP) + SMB (INV))

* HML: The return of cheap minus expensive stocks (value). Measures the difference between the returns on diversified portfolios of stocks with high and low Book-to-Market ratios.

HML = 1/2 (Small Value + Big Value) - 1/2 (Small Growth + Big Growth)

* RMW: The return spread of the most profitable firms minus the least profitable

(profit).

RMW = 1/2 (Small Robust + Big Robust) - 1/2 (Small Weak + Big Weak)

* CMA: The return spread of firms that invest conservatively minus aggressively (investment).

CMA = 1/2 (Small Conservative + Big Conservative) - 1/2 (Small Aggressive + Big Aggressive)

# Preprocessing

Data Cleaning

The first step in preprocessing is data cleaning. This involves checking for missing or NaN values in the data. Since the data is sourced from a reliable financial platform, it’s expected to be clean and reliable. However, non-trading days (like weekends and holidays) will not have any trading data, and these dates will not be included in the DataFrame.

Calculating Returns

The next step is to calculate the daily returns of the stocks. This is done using the logarithmic return formula:

The resulting DataFrame, returns, contain the daily logarithmic returns of the stocks instead of the prices. We have applied the log transformation, since the return data follows a Log normal distribution, to follow the basic assumptions of Linear regression. The first row of returns will be NaN because there’s no previous day’s price for the first day.

# Model Fitting

This analysis aims to understand the relationship between Amazon's (AMZN) returns and those of other significant stocks. Several models are built for this purpose: a simple Capital Asset Pricing Model (CAPM) and four multiple linear regression models with different sets of regressors.

Simple CAPM Model

The Capital Asset Pricing Model (CAPM), explains the relationship between systematic risk and expected return for assets, such as stocks. It is used for pricing risky assets, generating the expected return of an asset given its riskiness, and calculating the cost of capital.

The CAPM posits that the return of a risky asset is explained by the market factor. Mathematically, this can be expressed as:

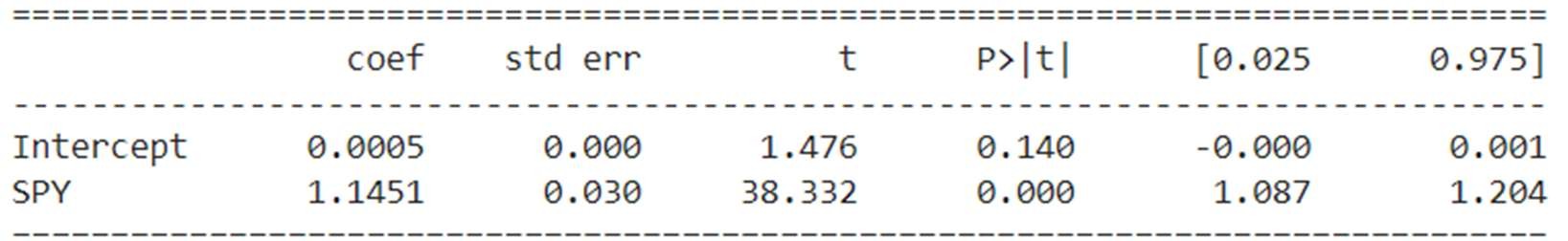
ra−rf=α+βM(MKT−rf)

Here’s what each term represents:

* **ra**: The return of the asset. This can be the return of any stock (e.g., Apple, Google, Tesla) or investment portfolio (e.g., any mutual/hedge fund portfolio).
* **rf**: The return of the risk-free asset. The US 3-month Treasury bill usually represents the risk-free asset, which is widely assumed in finance to be risk-free.
* **MKT**: The return of the market. This is usually represented by the S&P500 return, as it is the largest market index in the world.
* **α and βM**: The intercept and slope of the regression, respectively. βM measures the sensitivity of the asset’s returns to the market returns.

In the context of this analysis, a simple CAPM model is built using the ols() function from the statsmodels library in Python. The formula ‘AMZN ~ SPY’ is used, specifying that the returns of Amazon (AMZN) are the dependent variable and the returns of SPY (S&P 500 ETF) are the independent variable.

Following are the results obtained.



Adjusted R-Squared: **0.394**

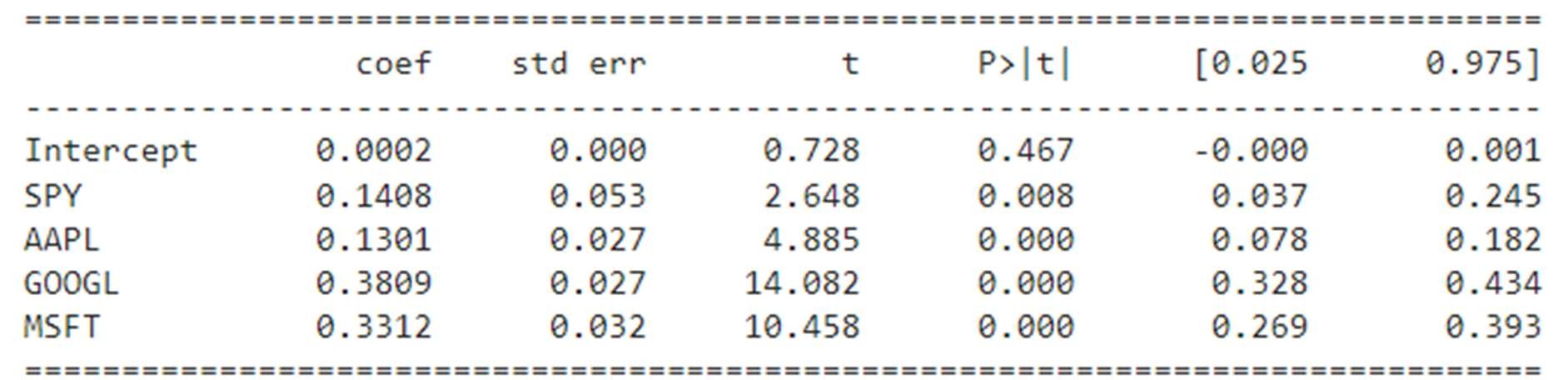
Prob (F-statistic): **4.03e-248**

Multiple Linear Regression Models

### 1. Stock Returns

The second model is a multiple linear regression model that includes four independent variables: the log returns of SPY, AAPL (Apple Inc.), GOOGL (Alphabet Inc.), and MSFT (Microsoft Corporation). This model aims to explain AMZN 's returns based on the returns of these four stocks.

Below are the results of the corresponding fit.

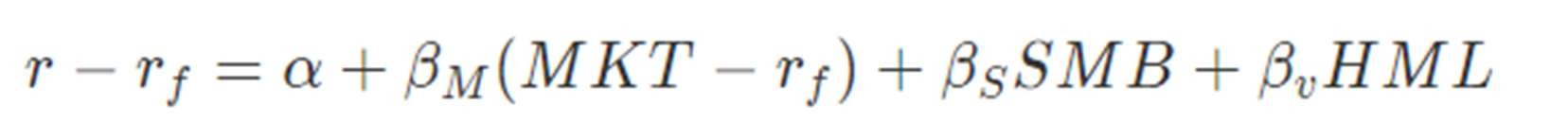


Adjusted R-Squared: 0.511

Prob (F-statistic): 0.00

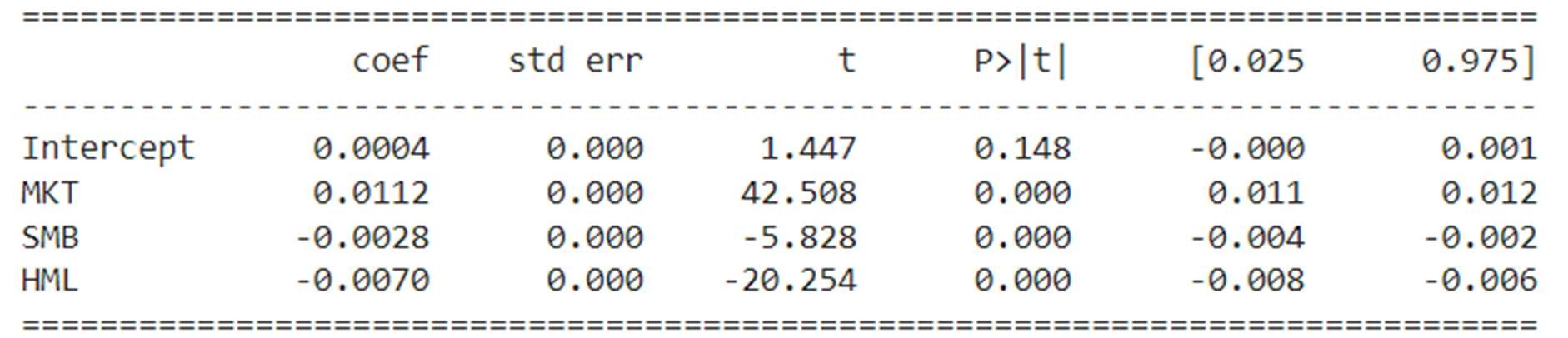
### 2. Fama-French 3-factor Model

Another very popular asset pricing model in the empirical finance literature is the Fama-French 3-factor (FF3) that was published in 1993. Nobel Laureate Eugene Fama and researcher Kenneth French found that value stocks tend to outperform growth stocks (i.e., value), and that small-cap stocks outperform large-cap stocks (i.e., size). Thus, the FF3 mode adds in size and value as risk factors to the model as shown below.



Where MKT, SMB and HML are the factors related to risk as explained in the Dataset section.

The results of the model were as follows:



Adjusted R-Squared: 0.519

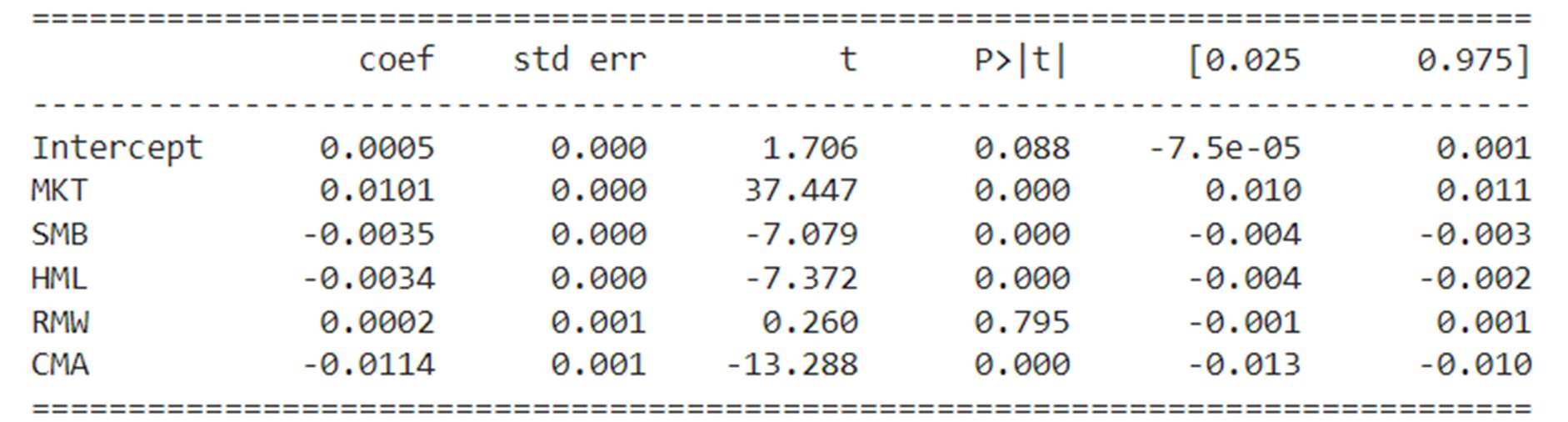
Prob (F-statistic): 0.00

### 3. Fama-French 5-factor model

In 2015, Fama-French added two more risk factors into their popular 3-factor asset pricing model to make a Fama-French 5-factor (FF5) model. This model added two 'quality' factors, namely profitability (stocks with a high operating profitability perform better) and investment (stocks of companies with high total asset growth have below average returns) factors and is modelled as below:



The results were as follows:

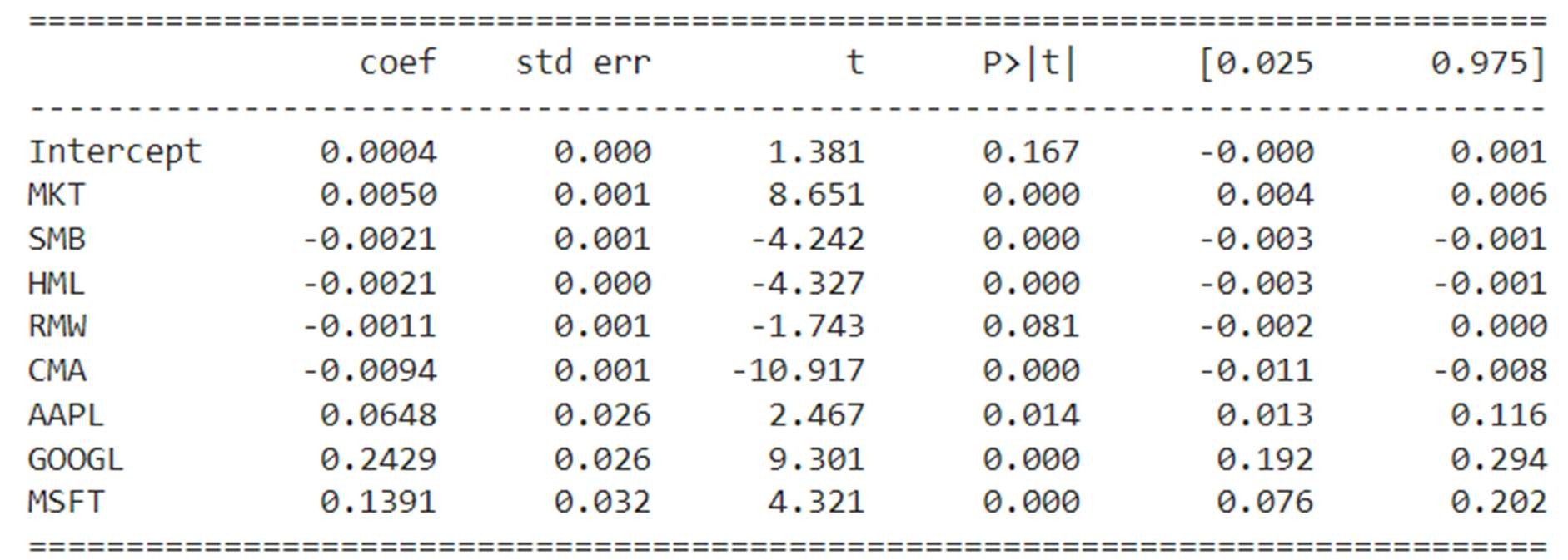


Adjusted R-Squared: 0.554

Prob (F-statistic): 0.00

4. Combined Model

The final model is a combined model that includes all the stock returns and FamaFrench factors. This model aims to explain the returns of AMZN by the returns of the other stocks (AAPL, GOOGL, MSFT) and the Fama-French factors (MKT, SMB, HML, RMW, CMA). The results were as follows:



Adjusted R-Squared: **0.578**

Prob (F-statistic): **0.00**

# Model Diagnostics:

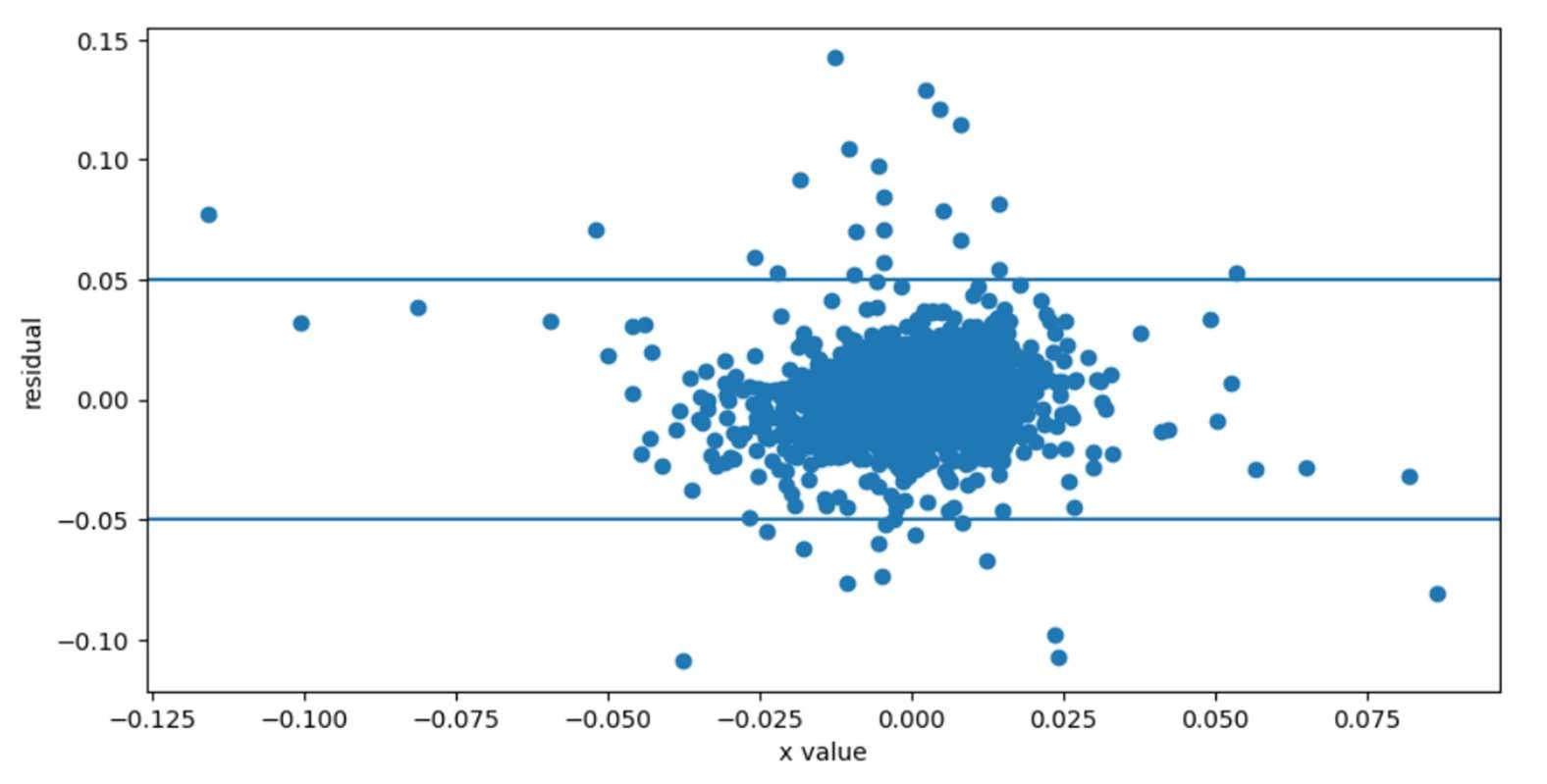
1. Residual Analysis

The residuals of the fitted model were plotted for each model. In all cases, through visual inspection, it was found that the density plot of the residuals followed an approximately normal distribution. Since normality of residuals is an assumption of the linear regression model, in our case, the assumption is proved to be valid and therefore, all inferences made on top of it should also be valid.

2. Homoscedasticity

In the case of the CAPM model, since there is only one regressor, homoscedasticity can be evaluated using a scatter plot. From the plot it can be observed that there is some change in the variance of the residuals when moving along the independent variable. The difference isn’t straightforward and could also be because there are more residuals in the middle than at the extremes.

In addition to the scatter plot, two horizontal lines are drawn at 0.05 and -0.05. These lines can serve as a reference to assess the dispersion of the residuals.



Since the plot doesn’t give an obvious answer, a hypothesis test can be used to confirm.

Breusch-Pagan test

We used the Breusch-Pagan test to check for heteroscedasticity. The Breusch-Pagan test is performed on the residuals of each model. It is a statistical test that checks if there is a significant relationship between the squared residuals and the independent variables. The null hypothesis of the test is that the residuals are distributed with equal variance (Homoscedasticity). A very low p-value therefore is indicative of heteroscedasticity.

Following are the p-values for each of the models fitted:

1. Simple CAPM Model: 0.0534
2. MLR with Stock Returns as Regressors: 2.3e-06
3. Fama-French 3-Factor Model: 0.291
4. Fama-French 5-Factor Model: 0.0012
5. Combined Model: 1.1990e-10

As can be observed from the p-values, CAPM and Fama-French 3-factor models are homoscedastic in nature. Therefore, the OLS method would be a valid model and would give statistically significant estimates for any inference made on top of it. In the remaining models, the presence of heteroscedasticity implies that the ordinary least squares (OLS) estimator of the covariance matrix of the regression coefficients may be biased, leading to incorrect standard errors and hypothesis tests.

This could be solved by using methods such as White's heteroscedasticity-consistent estimator (which assumes zero autocorrelation in the residual), which computes a robust covariance matrix for the regressor coefficients using the covariance matrix of residuals estimated using the error calculated from the model.

# Training Predictions

Visualising the last 20 return points of the training data predicted by each model and the actual returns.

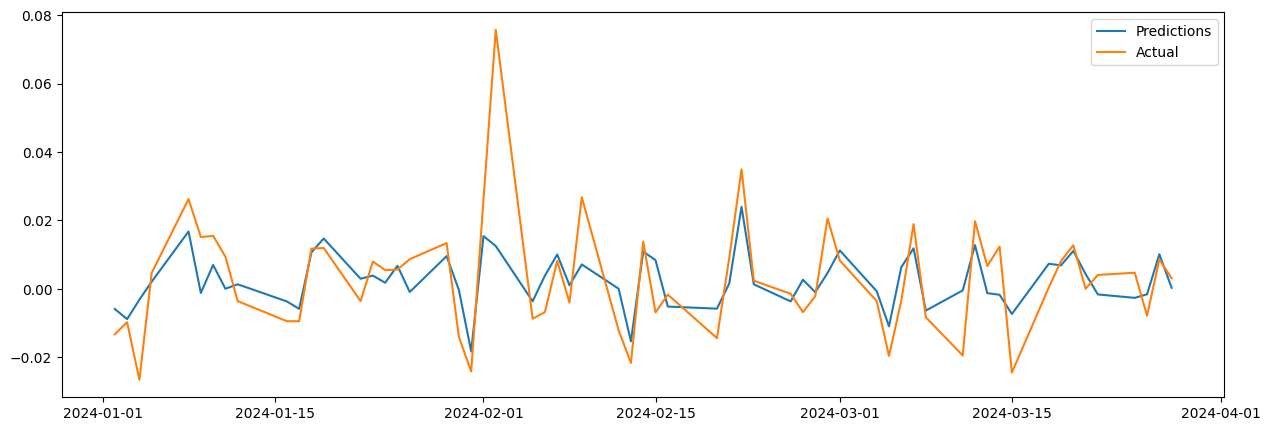


# Testing

Each model is then evaluated using the test data. For testing, stock returns and Fama-French factors from Jan 1st, 2024 to the end of February, 2024 were used. The Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE) were used to measure the accuracy of the predictions against the actual returns.

The results were as follows:

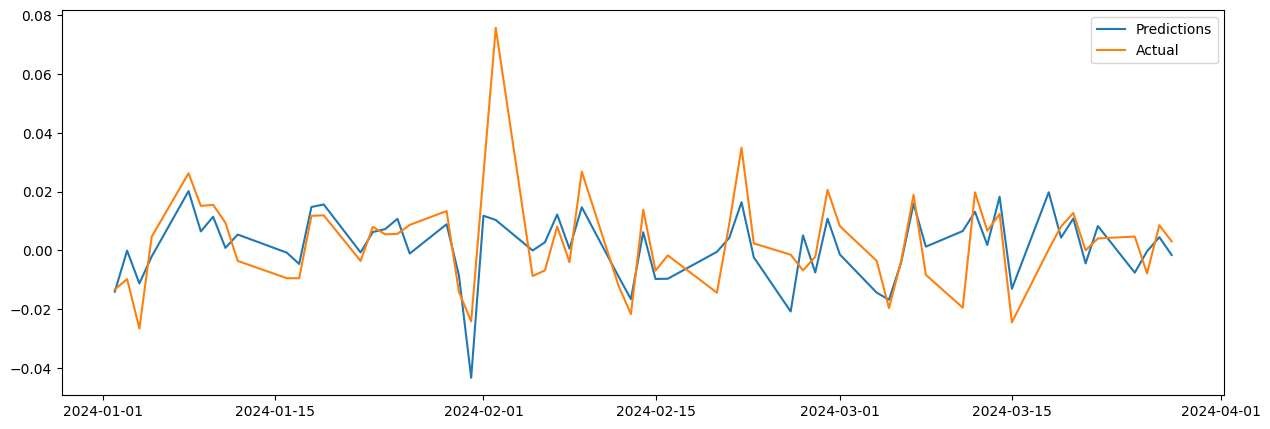
1. Simple CAPM Model



RMSE: 0.0119

MAE: 0.0081

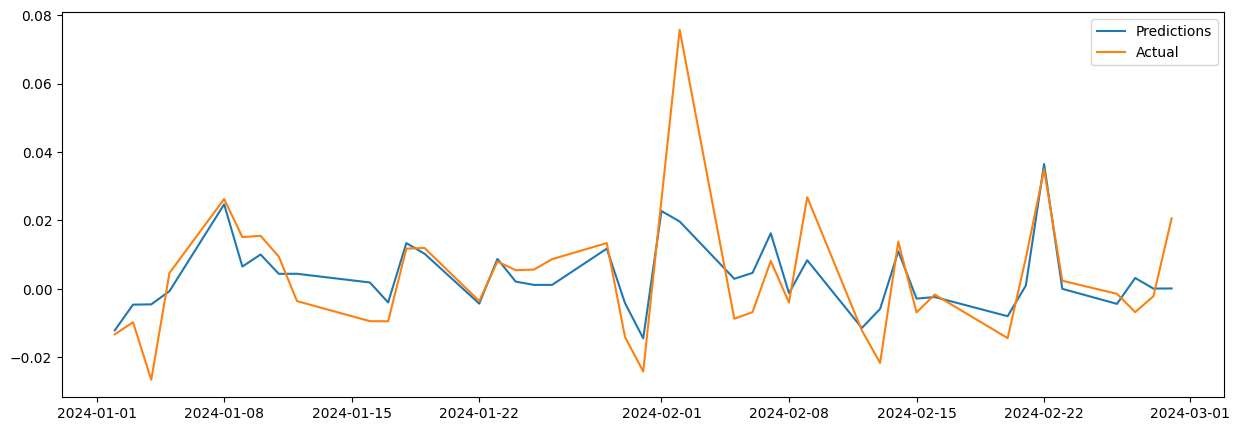
1. Multiple Linear Regression Model with Stock Returns as Regressors



RMSE: 0.0124

MAE: 0.0086

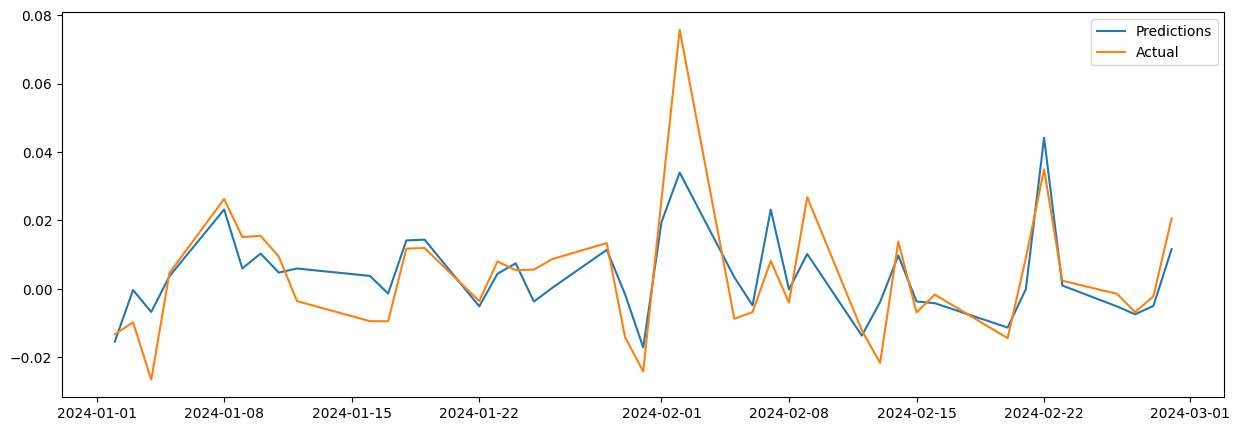
1. Fama-French 3-Factor Model



RMSE: 0.012

MAE: 0.0075

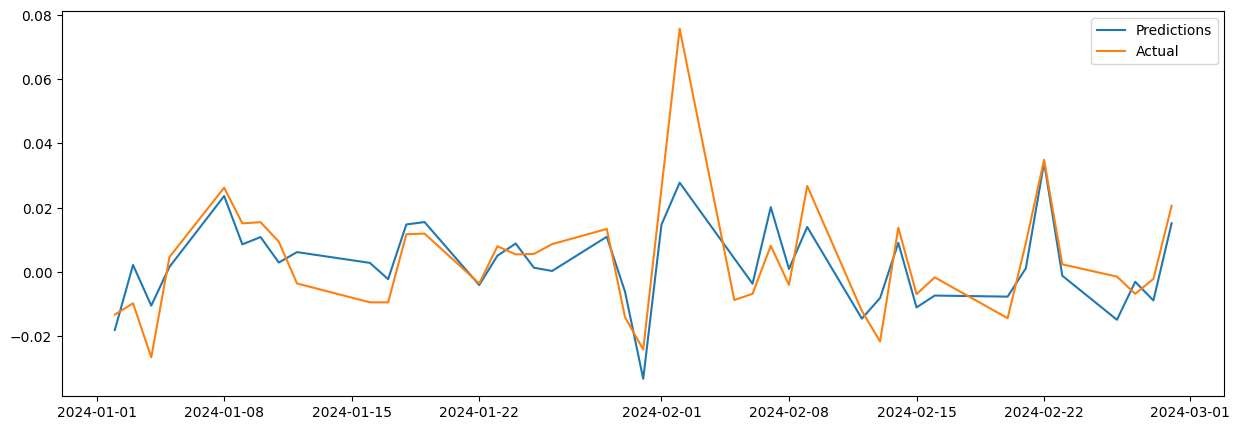
1. Fama-French 5-Factor Model



RMSE: 0.0104

MAE: 0.0073

1. Combined Model with all Stock Returns and all Fama-French Factors



RMSE: 0.0106

MAE: 0.0075

# Conclusion

The performance of all the models can be summarised as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Adjusted R-Squared | Prob (F-  Statistic) | p-value of  BreuschPagan  test | RMSE  (Testing) | MAE  (Testing) | Durbin  Watson  Coefficient |
| CAPM | 0.394 | 0.00 | 0.0534 | 0.0119 | 0.0081 | 1.961 |
| MLR with  Stock Returns | 0.511 | 0.00 | 2.3e-06 | 0.0124 | 0.0086 | 2.002 |
| Fama-French  3-factor | 0.519 | 0.00 | 0.291 | 0.012 | 0.0075 | 2.006 |
| Fama-French  5-factor | 0.554 | 0.00 | 0.0012 | 0.0104 | 0.0073 | 2.004 |
| Combined | 0.578 | 0.00 | 1.1990e-  10 | 0.0106 | 0.0075 | 2.008 |

Of all the models, the combined model has the highest Adj. R-Squared value, or in other terms, has captured the maximum variance in the data. But this model is heteroscedastic in nature and is therefore not a good fit for a linear regression model. The same is true for the FamaFrench 5-factor model and MLR model with stock returns as regressors.

These models could be a better option if we choose estimators which take into account the heteroscedastic nature of the residuals. The Durbin-Watson coefficients, as shown in the table above, for these models, are significantly close to 2, which indicates that there are no autocorrelations between the residuals. Therefore, as mentioned above, White's heteroscedasticity-consistent estimators can be used for these models so that their predictions are consistent.

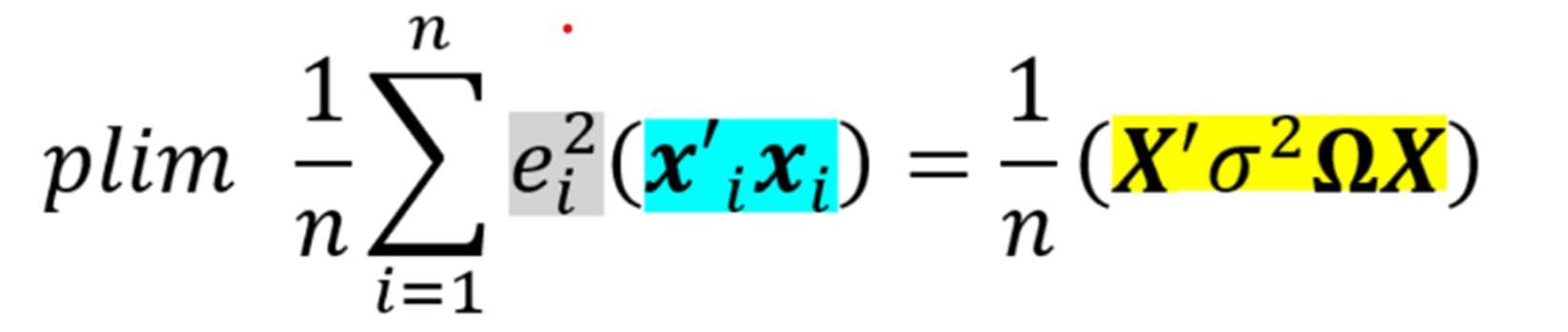
The following formula is the updated formula for the covariance matrix of estimators.

Var(B̂ ) = (X'X)^-1 X' σ² Ω X (X'X)^-1

where:

* X'X is the transpose of the design matrix X multiplied by itself
* Ω is the diagonal matrix of the true variances of the error terms
* σ² represents the unknown true variance of the error terms (often denoted by epsilon squared, ε²)

Since Ω and the variance factor σ² are unobservable as they relate to the error term ϵ of the regression model, we need to estimate them using the following the White’s formula.



Where 𝑒 is the error term.

Of the models which show constant variance in residuals, the Fama-French 3-factor model captures the maximum variance in the data. Both models perform similarly in the testing data with RMSE and MAE very close to each other.

Since all these models have residuals following a normal distribution, we can conclude that the Fama-French 3-factor model is the best fit among all the candidates we experimented with.