DS3000 Final Project

Ishaan, Deepanshu, Dev, Ignacio

DS 3000: Foundations of Data Science

Professor Mohit Singhal

December 8th, 2025

Abstract

This project studies how financial market variables relate to stock price movements and whether simple statistical models can explain and predict closing prices over time. Using daily price data from Yahoo Finance, we created engineered features such as daily percent change and moving averages and then applied multiple forms of regression. We compared linear, polynomial, and interaction based models to understand how different predictors influence accuracy. Results show that interaction based models capture more complex relationships between variables and can achieve strong in sample performance, while simpler models do not always generalize well. These findings suggest that single factor explanations of price changes are limited and that capturing interactions among financial variables is important for improving prediction.

Introduction

Understanding why a company's stock price moves is a central question in finance. While changes in revenue, profitability, and broader market conditions are commonly discussed, short run price behavior often depends on more immediate daily factors including volatility, trading volume, and momentum. Analysts and investors are increasingly interested in whether basic quantitative indicators can be used to explain price movements without relying entirely on subjective interpretation.

In this project, we explore whether observable market variables can help model daily closing prices using linear and polynomial regression methods. Our goal is not to build a perfect forecasting tool but to examine which features matter most and whether richer interactions between indicators improve model performance. This approach provides a foundation for future work that could include sector specific analysis or integration with fundamental financial statement data.

Data Description

We pulled daily historical stock price data using the Yahoo Finance Python interface. For each trading day, the dataset includes open price, high price, low price, close price, and trading volume. After downloading the raw data, we cleaned and transformed it by removing missing values and converting dates into a usable format. We engineered additional variables including daily percent change and two moving average indicators that capture short term and medium term trends in price behavior.

These processed features form the inputs to our regression models. Closing price serves as the response variable while open price, daily percent change, trading volume, and moving averages are used as explanatory variables. By comparing linear and polynomial specifications, we

evaluate whether interactions between variables such as price and volume provide additional information beyond simple linear relationships.

Method

The project applies two learning approaches, linear regression and polynomial regression. This was implemented directly in NumPy, which helped us ensure full control over the mathematical pipeline. Our goal is to model outcomes of stock market data. The regression-based methods provide an interpretable baseline that allows us to diagnose relationships between features and evaluate model assumptions.

The linear regression model method helps minimize the total squared error between predicted and actual values. Computing the coefficients using explicit matrix algebra. This helps inspect the model behavior directly. We felt that the linear regression was reasonable for this data because when needing to analyze the datasets, linear regression is a good starting point. This is a good starting point because the stock-price data set has many predictors, which often help show linear relationships over short windows. Along with that, it helps show interpretable coefficients, which help us understand how features predict the closing price for both analysts and domain experts. Even though financial markets are oftentimes nonlinear, having a linear regression remains the standard first model because it is mathematically clear, computationally efficient, and highly interpretable.

The polynomial regression was used because of the daily percentage change. This is because it may have nonlinear relationships with the closing price. Using polynomial regression helps expand a single feature into nonlinear transformations. Using the PolynomialFeatures(degree=4) converts the Daily-Change into x, x^2, x^3, x^4. Then it generates a matrix, which we then apply the ordinary least squares (OLS) estimator.

Linear and polynomial regression rely on several assumptions, which are linearity, independence of errors, homoscedasticity, normality of errors, and multicollinearity. For linearity, the OLS assumes a linear combination of predictors. The Polynomial model assumes that a fourth-degree polynomial is sufficient. For the independence of errors, there may be autocorrelation, which can lead to underestimated uncertainty. Homoscedasticity is the assumption that the OLS model assumes constant error variance. For the normality of errors, the residuals follow a normal distribution, which can make conclusions less reliable. Multicollinearity, which is model 1, assumes that the averages are strongly correlated. This can produce numerical warnings or coefficient explosions.

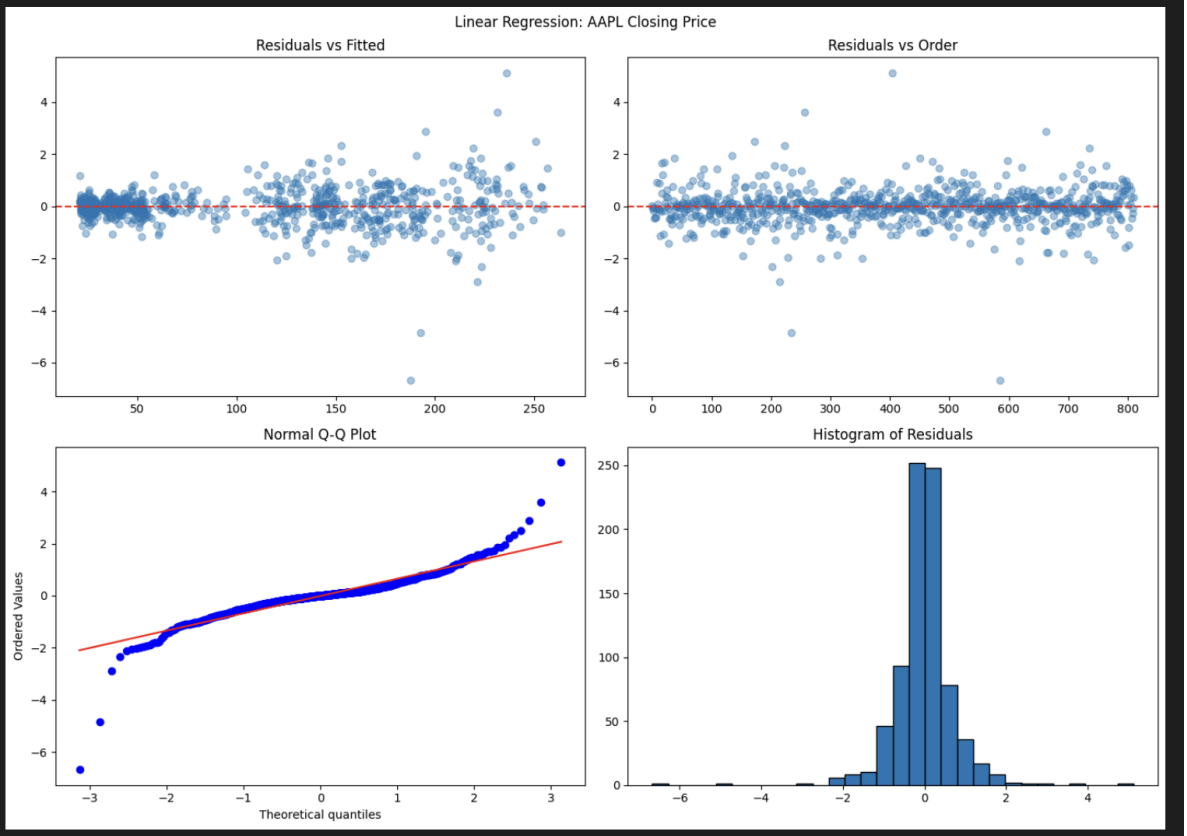
These methods are appropriate for the problem because of the objective of this project. The project is made for technical experts, where OLS helps provide mathematical results, Numpy,

which helps with computations and residual plots, MSE, and R^2 allow for assumption checking. It is also useful for application experts, where the linear model gives straightforward answers to financial indicators related to the closing price. The polynomial model is used to see if signal features show nonlinear effects.

Results

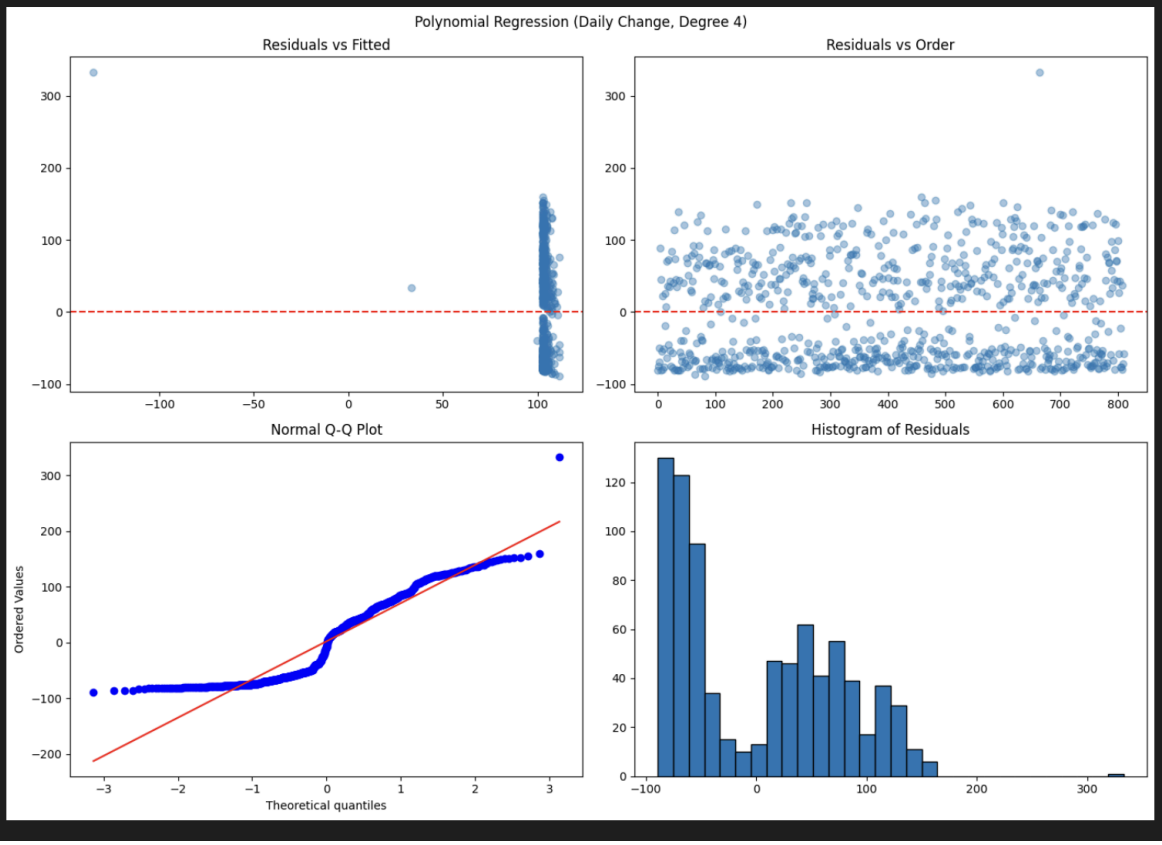
The linear regression model achieved strong predictive accuracy with an MSE of 0.500 and an R^2 of 0.9999, indicating that the model explains nearly all the variance in AAPL's closing price on the test set. Examination of the residuals versus fitted values plot shows that residuals are tightly clustered around zero with no clear curvature or funnel shape, suggesting that the assumptions of linearity and constant variance are reasonably satisfied. The residuals versus order plot displays residuals randomly scattered over time, with no obvious trends or cycles, indicating that the errors are largely independent and that the model does not systematically under- or over-predict during specific periods. The normal Q–Q plot shows residuals closely aligned with the theoretical normal line through most of the distribution, with only slight deviations at the extreme tails. This suggests that residuals are approximately normally distributed. Finally, the histogram of residuals is centered near zero and appears roughly symmetric, providing further evidence that the linear regression model produces well-behaved errors and a stable fit.

```
Linear Regression: AAPL Closing Price
MSE: 0.500
R²: 0.9999
```



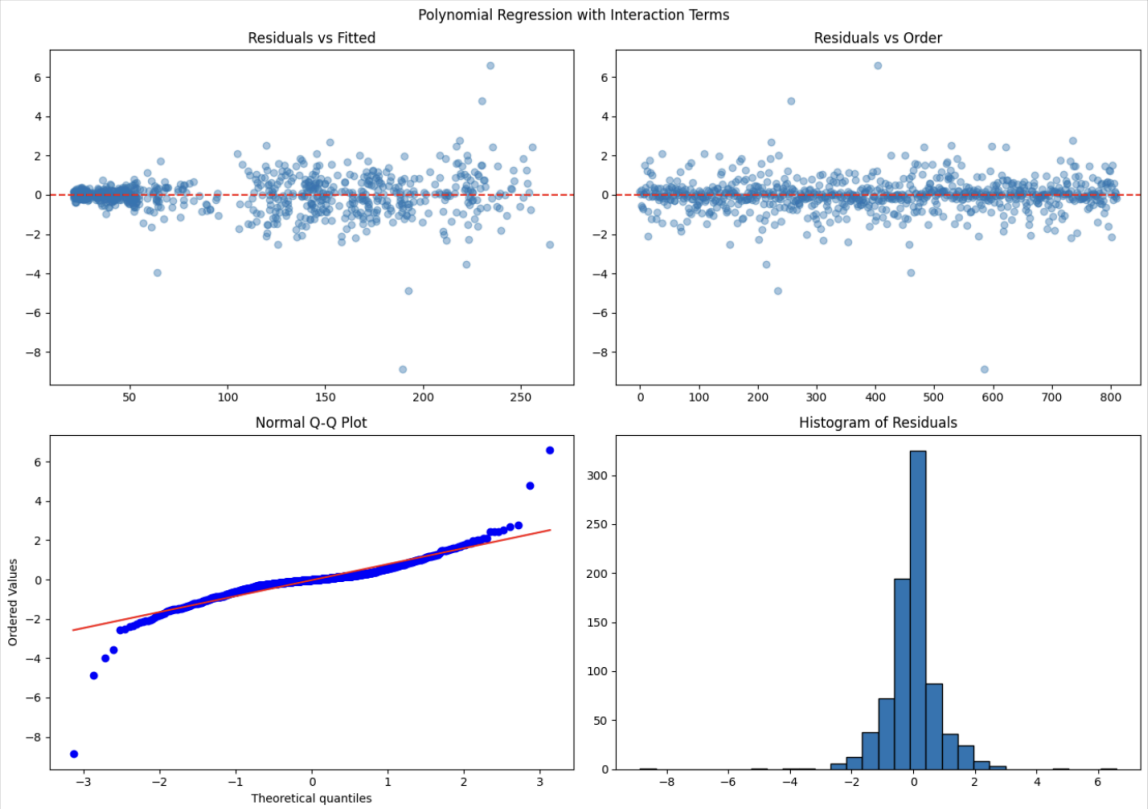
In contrast, the polynomial regression model using daily change (degree 4) performed poorly, with an MSE of 5224.4 and a negative R^2 value of −0.0249. The residuals versus fitted plot reveals extremely large residual magnitudes and clear vertical clustering, indicating that the model struggles to produce accurate predictions across the range of fitted values. This pattern suggests that the polynomial transformation of daily percentage change alone does not capture the underlying structure of the closing price. The residuals versus order plot further highlights instability, showing large blocks of consistently positive or negative residuals over time, which indicates significant temporal dependence and poor generalization. The normal Q–Q plot exhibits substantial departures from the reference line across nearly all quantiles, with heavy tails and sharp curvature, demonstrating severe violations of the normality assumption. Additionally, the histogram of residuals is highly skewed and multimodal rather than centered around zero, confirming the presence of extreme prediction errors. Overall, all four diagnostic plots consistently indicate that the daily-change-only polynomial model is poorly specified and unreliable.

```
Polynomial Regression (Daily Change, Degree 4)
MSE: 5224.400
R²: -0.0249
```



The polynomial regression model with interaction terms shows a substantial improvement over the simpler polynomial model, achieving an MSE of 0.800 and an R^2 of 0.9998, closely approaching the performance of the linear regression model. In the residuals versus fitted values plot, residuals are again centered around zero with relatively constant spread, indicating that the inclusion of interaction terms helps account for nonlinear relationships without introducing major heteroskedasticity. The residuals versus order plot shows a largely random scatter of points across time, suggesting that the model does not suffer from strong temporal dependence in its errors. While the normal Q–Q plot shows minor deviations at the extreme tails, most residuals align well with the theoretical normal distribution, indicating improved normality compared to the daily-change polynomial model. The histogram of residuals is sharply centered around zero and approximately symmetric, with far fewer extreme values. Together, these diagnostic plots indicate that incorporating interaction terms between market variables significantly stabilizes the polynomial regression and enables it to capture meaningful structure in the data without excessive error variance.

```
Polynomial Regression with Interaction Terms
MSE: 0.800
R²: 0.9998
```



Discussion

The results produced from our models show how predicting closing stock prices can introduce problems even when we use tested regression techniques. The linear regression produces the most accurate results as we have already discussed. However, it fails to highlight how dependent the closing price is to movements in the stock market as a whole and other financial indicators. Even though our model produces very impressive results, this does not necessarily imply that there is a clear relationship. Instead, it shows how tightly linked variables like open, high and low already are in the stock market as well as how little movement there is between these metrics throughout the day.

We then looked at the polynomial regression that only relied on daily percent change, which performed significantly worse than our previous model. The daily percent change of stocks fluctuates significantly over a longer period of time and this did not allow the model to capture how this related to closing price. From our results in the residual patterns and the extreme errors,

we can see that the model was trying to predict a stable outcome using a signal that was very unstable. Even though the daily percent change might be a reasonably good indicator of volatility in the short term and changes in momentum, it is not effective at determining the closing price of a stock by itself. As a result, the polynomial transformation ended up amplifying noise rather than uncovering any meaningful relationship.

On the other hand, the polynomial model that also included interaction terms had a significantly better performance. We included more features like open, high, volume, etc. which gave the model more information to work with, helping it recognise some of the patterns amongst the stock market data. Given that the accuracy of this model improved greatly and had more controlled residual behaviour, it can be said that the features used and their interactions play a significant role in determining the closing price for a given stock. However, there were still some limitations and the model did not outperform our original linear regression model by a significant margin. This suggests that the added complexity only helps when it captures genuine structure in the data, and in this case, most of that structure was already being handled well by the linear model.

In terms of our original question of whether basic market indicators can be used to model closing prices, our results suggest that we did find somewhat of a solution. Both the linear and the polynomial models show that variables like open, high, low and moving averages produce an effective reconstruction of the closing price. Also, the values for R^2 were extremely high and the errors very low made us question whether we were actually predicting the closing price or if we were just expressing values that were already very close and related to the close price. In addition, our residual plots still assume independence in time and constant variance, and those assumptions may not fully hold in real financial markets, where individual large cap companies can cause movements in the market. This made us realise that our model might not have been as effective as we once believed. It is more of a tool to understand how the indicators we used move together rather than aiding investment decisions.

In conclusion, our findings show that predicting the daily close price for a stock is more complicated than it may seem at first. Even though the metrics we used appeared to be strong, there are underlying relationships between features that are influenced by many factors that our feature set did not include. The failure of the daily change model and the modest improvement from interaction terms both highlight how important it is to use features that reflect real financial behavior. If we were to extend this work, adding time-series methods, additional market indicators, or more flexible modeling techniques would likely give a fuller picture. But within the scope of this project, linear regression remains the most reliable and interpretable model for understanding how our chosen variables relate to the closing price.