STAT 27410 Final Project Proposal - A Bayesian Approach to Portfolio Management

Andrew Su, Scarlett He

1. Introduction

Portfolio management is a complex endeavor that involves not only selecting appropriate investments but also navigating the uncertainties posed by market dynamics and economic conditions. While diversification across sectors and asset classes is widely regarded as a cornerstone of effective portfolio construction, managing risk and optimizing returns remains a significant challenge. This study seeks to explore innovative approaches to enhance investment decision-making by comparing traditional frequentist methods with Bayesian techniques in the context of portfolio optimization.

Over a three-year period from January 1st, 2022, to January 1st, 2025, we analyze historical stock prices for five individual stocks:

- Apple Inc. Common Stock (AAPL) (Nasdaq, n.d.b)
- Coca-Cola Company (The) Common Stock (KO) (Nasdaq, n.d.c)
- Costco Wholesale Corporation Common Stock (COST) (Nasdaq, n.d.d)
- Advanced Micro Devices Inc. Common Stock (AMD) (Nasdaq, n.d.a)
- Salesforce Inc. Common Stock (CRM) (Nasdaq, n.d.f)

Two sector specific ETFS:

- Invesco QQQ Trust (QQQ) (Nasdaq, n.d.e)
- the Consumer Discretionary SPDR Select Sector Fund (XLY) (Nasdaq, n.d.h)

As well as S&P 500 index (SPX) (Nasdaq, n.d.g) as a baseline for comparison.

All data used is from the 3-year timespan of January 1st, 2022 to January 1st, 2025 (753 trading days) and obtained directly from the Nasdaq stock exchange. Data was sourced directly from the Nasdaq stock exchange, ensuring a robust empirical basis for our analysis.

The selection of these securities reflects a strategic focus on diverse industries and market segments. We chose individual companies and ETFs to represent both the Technology and Consumer sectors, ensuring a mix of defensive and cyclical stocks while avoiding undue

similarity among the selected investments. This approach aims to capture variations in market performance and provide a comprehensive view of portfolio dynamics.

For each of the securities, separate datasets were used. For each security, the data set include the following variables:

- **Date**: The date of the trading day
- **Close/Last**: The price of the security at the end of the trading day
- **Volume**: The total number of shares traded during the trading day
- **Open**: The price of the security at the start of the trading day
- **High**: The highest price of the security during the entire trading day
- **Low**: The lowest price of the security during the entire trading day

For the purpose of our analysis, we will only be looking at the Date, Close/Last, Open and Volume categories across each of the securities to build a fundamental AR model to capture essential market dynamics and trends.

To determine weights of allocation, we will run regression models to predict the future prices of each stock, calculating the percentage changes of each and assigning portfolio weights through convex optimization. For instance, for a certain stock A, our model will include the previous prices of stock A into our calculation for future price of A, while also incorporating the sector-ETF and SPX as measures of how well the economy is doing.

We define the stock price of stock i at time t as $P_{i,t}$. The predictive model for stock prices is given as follows:

For each technology sector stock, the model is expressed as:

$$\begin{aligned} P_{i,t+1} = \beta_1 P_{i,t} + \beta_2 P_{i,t-1} + \\ \beta_3 \text{QQQ}_t + \beta_4 \text{SPY}_t + \sum_{j \neq i} \text{COV}(P_{i,t}, P_{j,t}) \end{aligned}$$

For each consumer sector stock, the model is expressed as:

$$\begin{aligned} P_{i,t+1} = \beta_1 P_{i,t} + \beta_2 P_{i,t-1} + \\ \beta_3 \text{XLY}_t + \beta_4 \text{SPX}_t + \sum_{j \neq i} \beta_{5i} \text{COV}(P_{i,t}, P_{j,t}) \end{aligned}$$

The percentage change in stock price from time t to $t+1$, denoted as $R_{i,t}$, is given by:

$$R_{i,t} = \frac{P_{i,t+1} - P_{i,t}}{P_{i,t}}$$

As opposed to a pure Auto-Regressive (AR) model, we will consider exogenous variables that aim to capture aspects of the economy that cannot be fully captured with our limited selection of stocks. This would allow for specific analysis while also factoring in more complex market trends. *Note: Our models are subject to change.*

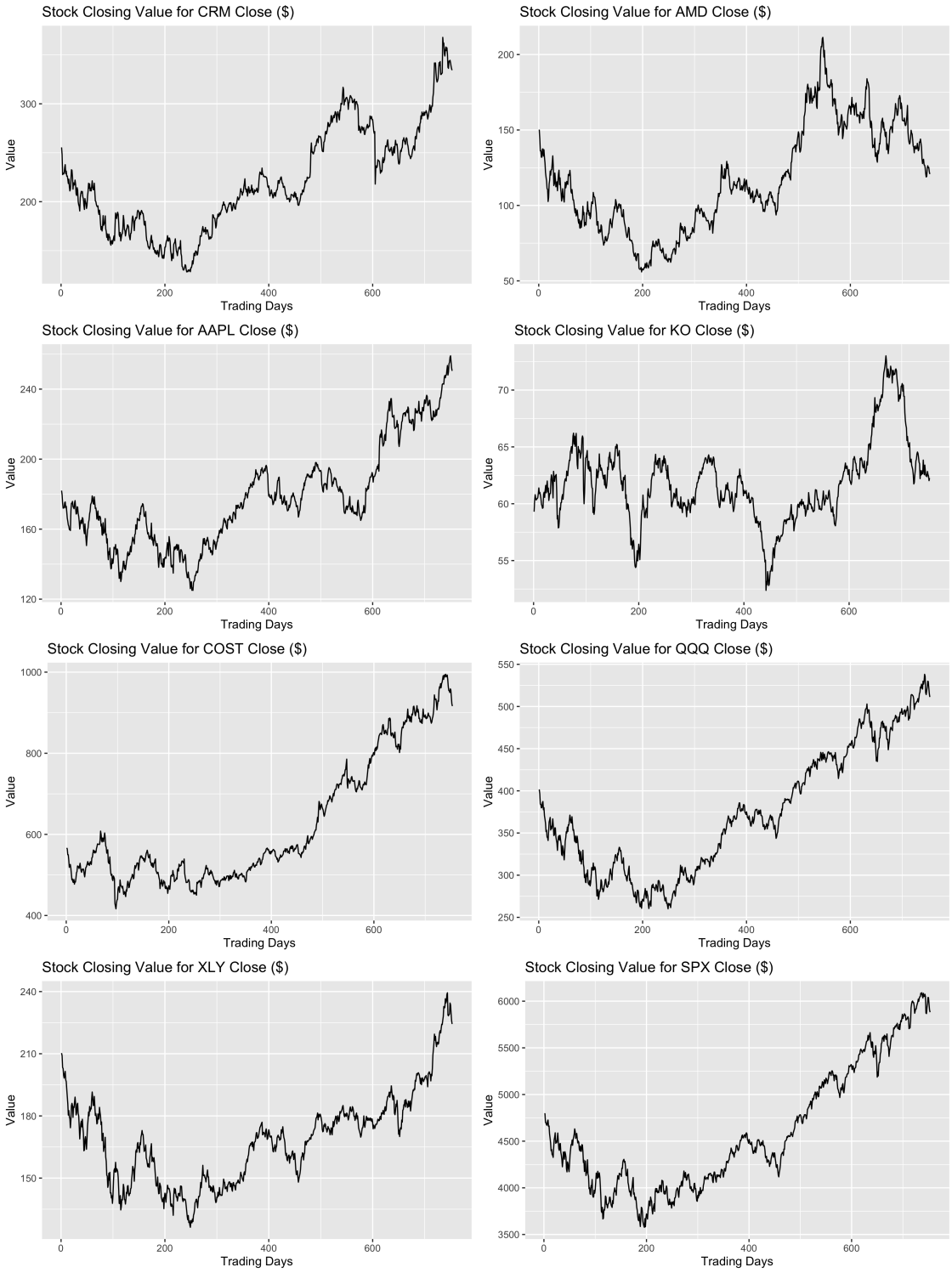
With computed stock price percentage changes, we will use convex optimization to determine the portfolio weights, where volume traded will play a role in managing the risk assigned to each stock.

We will then compare these frequentist approaches with Bayesian methods, where we will assume a prior distribution (multivariate normal or student-t) on an asset, and then continuously update the distribution using the market data as the likelihood function to generate a posterior distribution that gives the most likely expected return for assets and measuring uncertainty through credible intervals. These returns can then be used in combination with the uncertainty measures to allocate portfolio weights.

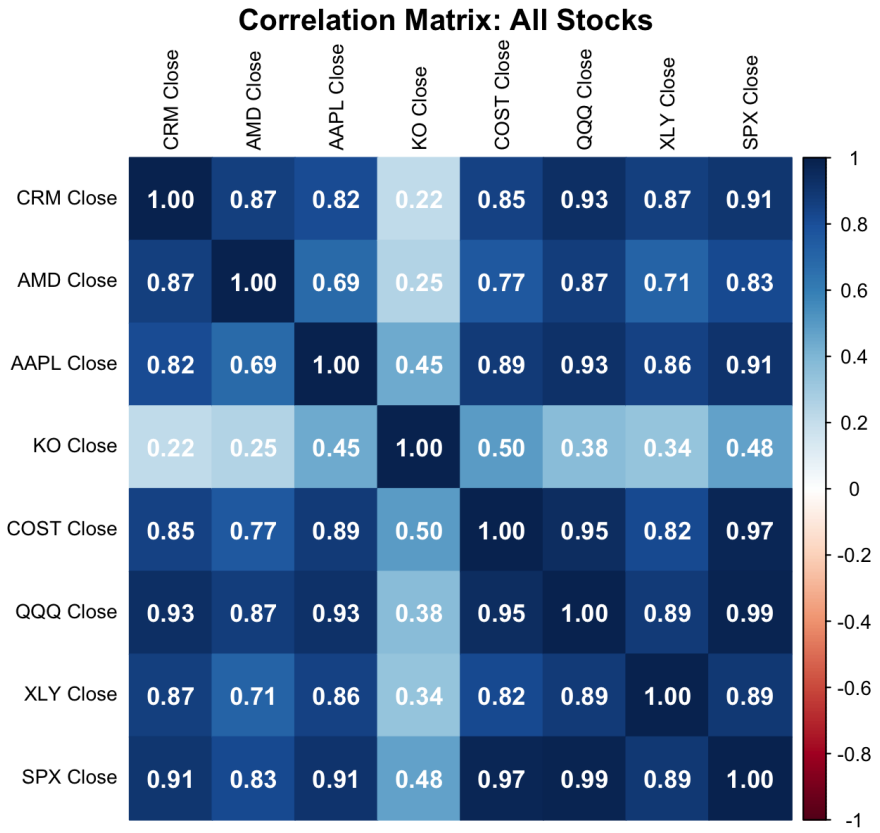
This study not only sets the stage for a detailed comparison between frequentist and Bayesian methods in portfolio optimization but also provides insights into the important features to capture comprehensive market dynamics.

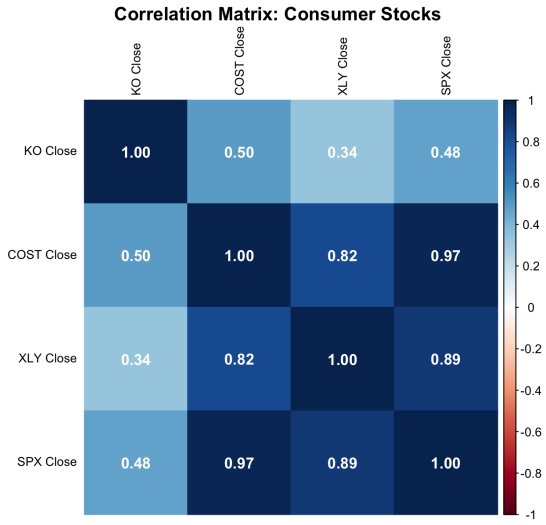
2. Exploratory Data Analysis

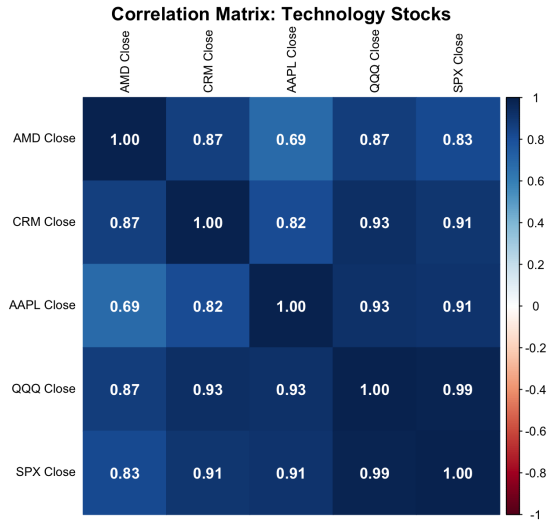
We will start by looking at the trends and changes in each of the securities over this 3 year period.



These graphs show the trends of the 7 securities over the 3-year period. It can be noted that some graphs have similar trends. Thus, we look at correlation matrix of those stock, and we found the correlation is relatively high due to the choice of our stocks are all well known and having a high market share.







3. Frequentist Analysis

3.1 Proposed Frequentist Model(s)

3.1.1 Time Series Models

Time series analysis emerges as a fundamental approach for financial data forecasting due to financial data's natrual of time dependency. It offers ways to locate patterns from historical fianncial data and form predictions to provide insights on investment decisions.

The strength of time series analysis lies in its capacity to clean data and remove confounding variables and white noises. It also handles the non-independency between data in a sequence of time.

In our methodological approach, the initial phase involved testing for stationarity in the financial time series data. TWe observed that the majority of financial time series exhibited non-stationary characteristics, which needs initial differencing for the data.We implemented first-order differencing, which shown successful transformation of the data into a stationary series, as confirmed by the Augmented Dickey-Fuller (ADF) test.

Following the establishment of stationarity, we examined Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. These diagnostic tools provided crucial insights into the underlying structure of the time series, particularly in determining the presence and order of autoregressive (AR) and moving average (MA) components. The patterns observed in these plots served as preliminary indicators for model specification.

To ensure optimal model selection, we fitted multiple ARIMA models to each financial time series, varying the orders of AR or MA each time, and compare them. Model performance was evaluated using established criteria such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion).

In the end, we conducted model diagnostic by checking the residuals. We ensured model residuals fell within acceptable bounds and exhibited properties consistent with white noise processes. This validation process confirmed the adequacy of our selected models and their suitability for forecasting purposes.

3.1.2 Multiple Linear Regress (MLR) Models

Linear regression serves as a fundamental forecasting method. Multiple Linear Regression extends the basic model by incorporating various market factors, enabling deeper analysis of market behaviors.

Our methodology implements recursive prediction, with the initial model expressed as:

$$P_{i,t+1} = \beta_1 P_{i,t} + \beta_2 P_{i,t-1} + \beta_3 \mathrm{QQQ}_t + \beta_4 \mathrm{SPX}_t$$

where $P_{i,t+1}$ represents the predicted price, $P_{i,t}$ and $P_{i,t-1}$ represent lagged prices, and QQQ_t, SPX_t represent market indices at time t.

Model validation employs multiple diagnostic tests. We examine R-squared values and adjusted R-squared to assess model fit, F-statistics for overall significance, and t-statistics for individual variable significance. Residual analysis confirms assumptions of normality, homoscedasticity, and independence.

The recursive prediction process follows these steps: 1. Predict future price using lagged values 2. Update the dataset with predicted values 3. Adjust lagged variables for subsequent predictions 4. Iterate for next time period prediction This recursive approach enables continuous updating of predictions while maintaining the temporal structure of the data.

3.2 Fitting the Frequentist Model(s)

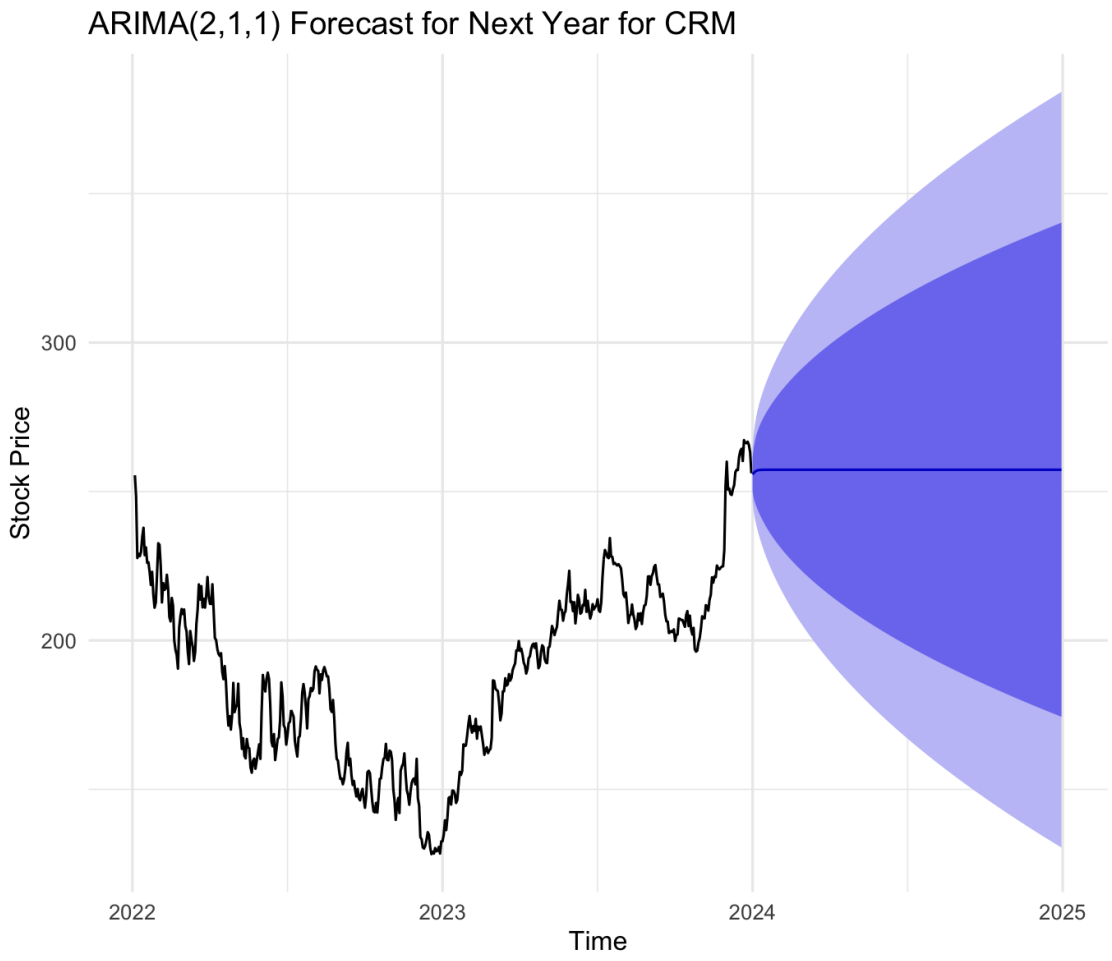
3.2.1 Fitting Time Series Models

We discovered that time series data is better for fitting short term future data. The variance of the prediction increases as long as forecasting period extends, limiting the ability to provide insightful prediction.

Using CRM stock as a case study, we identified ARIMA(2,1,1) as the optimal model specification through diagnostic testing. The model parameters indicate significant autoregressive and moving average components, suggesting the stock price exhibits both momentum and mean-reverting characteristics.

```
Call:
arima(x = CRM_ts, order = c(2, 1, 1))

Coefficients:
         ar1      ar2      ma1
      0.7637  -0.1119  -0.6912
s.e.  0.1883   0.0450   0.1852

sigma^2 estimated as 21.02:  log likelihood = -1473.77,  aic = 2955.54

Training set error measures:
                     ME    RMSE      MAE
Training set 0.00553772 4.57998 3.396958
                     MPE     MAPE      MASE
Training set -0.03189323 1.827166 0.9937438
                     ACF1
Training set -0.002146988
```

However, we discovered that mean of the prediction remains consistent, while the prediction intervals expand over time.



Similar trends exist in other stocks, which limit the possibility for us to perform future portfolio optimization.

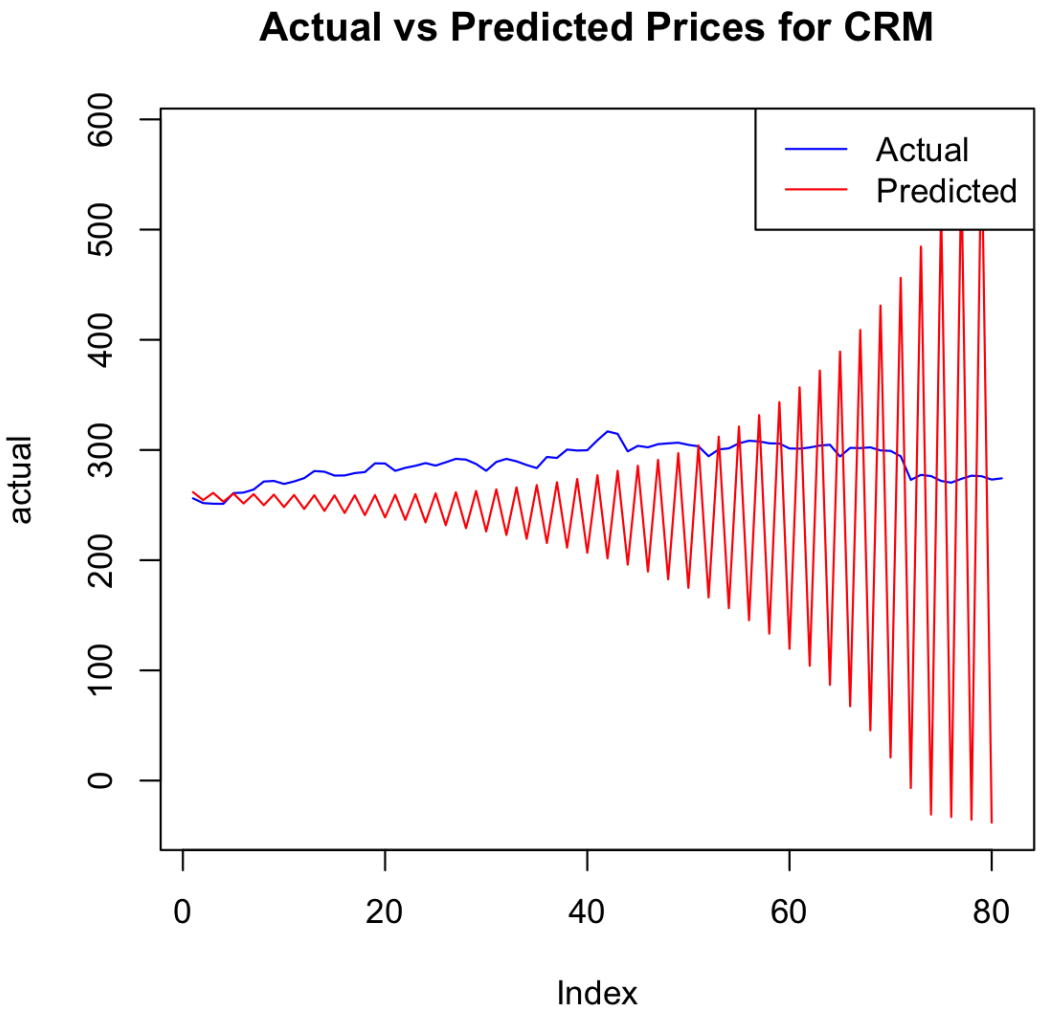
3.2.2 Fitting the MLR Models

Our Multiple Linear Regression model encounters same challenges in forecast accuracy. As predictions extend further from known data points, subsequent forecasts increasingly rely on previously predicted values rather than actual data. This dependency creates a compounding effect, leading to expanded variance and prediction intervals over time.

Analysis of the model coefficients, as shown in the regression summary, reveals that the most significant predictor variable is the immediate lagged price, with the highest t-value and lowest p-value. This finding aligns with market efficiency principles, suggesting that recent price information captures the most relevant market signals.

```
lm(formula = CRM_Close ~ CRM_lag1 + CRM_lag2 + QQQ_lag1 + SPX_lag1,
    data = stock_train)

Residuals:
     Min       1Q   Median       3Q      Max
-18.1038  -2.7298  -0.0585   2.7064  21.0013

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.108903   5.309049   3.411 0.000700
CRM_lag1     1.055479   0.046350  22.772  < 2e-16
CRM_lag2    -0.077980   0.044779  -1.741 0.082229
QQQ_lag1     0.065850   0.024643   2.672 0.007785
SPX_lag1    -0.008454   0.002547  -3.319 0.000971

(Intercept) ***
CRM_lag1    ***
CRM_lag2    .
QQQ_lag1    **
SPX_lag1    ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.553 on 495 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.9782,    Adjusted R-squared:  0.978
F-statistic:  5553 on 4 and 495 DF,  p-value: < 2.2e-16
```

However, the prediction is influenced by the fluctuations of stocks and is increased due to recursive prediction.



The visualization shows how the prediction intervals expand significantly as the forecast horizon extends, forming a characteristic "cone of uncertainty." This pattern is particularly evident in the latter portion of the forecast period, where the model's predictions exhibit increased oscillation and deviation from the central trend. This behavior underscores the limitations of recursive prediction in long-term forecasting applications, suggesting the model may be more reliable for shorter-term predictions where the cumulative impact of forecast errors remains contained.

4. Bayesian Analysis

Propose the Bayesian analysis you will work on during the rest of the quarter in this session.

4.1 Proposed Bayesian Model(s)

In this section,

- formulate the Bayesian model(s) you are going to use to analyze your dataset. Be sure to first define the notations involved in the model(s).
- discuss how you will elicit the prior(s).

4.2 Fitting the Bayesian model(s)

- Propose how you will fit the proposed Bayesian models.
- Propose how you will perform sensitivity analysis of the Bayesian models, i.e., how the posterior distribution is affected by the prior
- Propose how you will check the MCMC convergence.

4.3 Prediction

In this section, propose how you can make predictions using the Bayesian model.

5. Discussion

In this section, discuss how you can improve your model.

6. Contributions

In this section, discuss the percentage of your contributions to the development final project proposal. Report the number of hours you have worked on the proposal, and the sections you are involved.

Please also discuss briefly the contributions of your teammate(s), as well as the help and support you got from your teammates(s).

References

Nasdaq. n.d.a. *Advanced Micro Devices, Inc. Common Stock (AMD) Historical Quotes*. https://www.nasdaq.com/market-activity/stocks/amd/historical.

———. n.d.b. *Apple Inc. Common Stock (AAPL) Historical Quotes*. https://www.nasdaq.com/market-activity/stocks/aapl/historical.

———. n.d.c. *Coca-Cola Company (the) Common Stock (KO) Historical Quotes*. https://www.nasdaq.com/market-activity/stocks/ko/historical.

———. n.d.d. *Costco Wholesale Corporation Common Stock (COST) Historical Quotes*. https://www.nasdaq.com/market-activity/stocks/cost/historical.

———. n.d.e. *Invesco QQQ Trust, Series 1 (QQQ) Historical*. https://www.nasdaq.com/market-activity/stocks/qqq/historical.

———. n.d.f. *Salesforce, Inc. Common Stock (CRM) Historical Quotes*. https://www.nasdaq.com/market-activity/stocks/crm/historical.

———. n.d.g. *S&p 500 (SPX) Historical Data*. https://www.nasdaq.com/market-activity/index/spx/historical.

———. n.d.h. *SPDR Select Sector Fund - Consumer Discretionary (XLY) Historical*. https://www.nasdaq.com/market-activity/stocks/xly/historical.

Appendix

https://raw.githubusercontent.com/asu1-1/STAT27410-Final-Project/refs/heads/main/27410%20Final%20Project%20Code.R

https://raw.githubusercontent.com/asu1-1/STAT27410-Final-Project/refs/heads/main/Scarlett_part1_code.R