Index Fund Price Replication from a Subset of its Sub-Indices: A Replication of Dow Jones Industrial Average Price from the Prices of a Subset of its Sub-Indices

Calvin Emmry Okello, Kyaw Lin Oo, Mohammad Ilyas Zewar

WorldQuant University

okellocalvinemmy@gmail.com, *klo.kyawlinoo99@gmail.com*, *m.ilyas.zewar@gmail.com*

A Project Proposal Submitted to WorldQuant University for the Partial Fulfillment of the Award of Master of Science in Financial Engineering

1 Background

1.1 Introduction

Index funds have become a benchmark that most investors try to optimize their portfolio to outperform. Some investors replicate the entire sub-indices and optimize the composition for a better performance, however, this is associated with very high trading fees. Alternative to full replication, some investors take index funds and select a subset of its sub-indices to have a better performance since the sub-indices that would lower the portfolio performance would be left out (Blume & Edelen, 2002). The selection of a subset of index fund sub-indices greatly reduces the fees associated with trading the replicated portfolio.

The selection of a subset of index sub-indices sounds appealing but without proper selection of the sub-indices, huge tracking error might arise. This misleads the investor and causes low returns or losses. Since the starting point of achieving a higher return would be to track the index fund itself, using the key drivers of the index fund can be an ideal method of selection of its key drivers. Key drivers push the index fund not only in terms of its direction of movement but also in terms of quantity. This makes Pearson correlation a good candidate for model selection since it considers directions and strength.

Selecting the key drivers to replicate the index fund may not be optimal enough to replicate the index fund but proper allocation of resources to the selected sub-indices is of paramount importance to make sure the actual replication is done. This optimization task can be quite challenging, however, with advancement in technology, the optimization tasks have been developed. A simpler way would involve training a regression model with sub-indices as predictors of the index fund. This is believed to properly allocate the proportions of distributions of resources that replicate the index fund with minimal tracking error.

The study is intended to develop a model that replicates an index fund (Dow Jones Industrial Average) from its 5 most correlated sub-indices. The study shall involve using a regression analysis with selected sub-indices as the predictor of the index fund. The study shall also utilize time series analysis to project out of the sample prices of the selected sub-indices which are in turn used to predict the index fund prices out of sampled period.

1.2 Problem Statement

“An index fund might not track the underlying index or sector exactly, causing tracking errors or variances between the fund and the index” (Croome). Moreover, indices are constantly changing, and “when an index changes, the movement can affect the price of individual stocks being added or dropped, as well as funds based on the index itself” (Charles Schwab). “Understanding the mechanics of index shake-ups, and how they can affect your portfolio, can help identify trading opportunities by ensuring you are not caught off guard when changes occur” (Charles Schwab). Investors who track the underlying indices of an index fund often face the challenge of losing their preferred securities when changes are made to the index. Some investors may consider including such stocks in their portfolio independently. Studying the underlying indices that predominantly correlate with the price movement of an index fund could be a more effective way of replicating an index fund, helping investors prepare for unexpected changes that might cause losses due to unpreparedness.

1.3 Goals and Objectives

- Determine the most likely replicas of an Index fund from its Subsets
- Analyze Time Series Characteristics and Performance of index fund compared to its subsets
- Develop a model replicating an index fund from subset of its sub-indices

2 Literature Review

Index Funds have become one of the investment portfolios that investors tend to invest in, majorly due to its self-diversification and being a managed fund which is carefully always monitored. Index funds have been found to be subject to changes most especially whenever a stock is removed and or added. Considering index fund replication prepares the investor to allocate his/her portfolio in a way that prepares him/her for the changes.

(Dyer and Guest) manually read the PIS in funds' 497Ks and found out that while the choice to replicate vs. sample appears to be fairly stable over time in general, it is possible that funds could change approaches over time. This makes the model not perform the way an investor would expect and thus, it is better for such investors to adopt a method and/or a model that continues to perform in an expected manner. Index replication is one such method to consider however, replications face some challenges for instance high trading fees.

Index replication is also faced with challenges that occur when a stock is announced to be added or deleted from an index fund. (Blume and Edelen) reviewed the earlier studies and found out that individual stocks realize, on average, positive abnormal returns from the date of announcement of adding the stock into an index fund through the change (addition) day. They also reviewed that stocks deleted from index funds realizes negative abnormal returns. This does affect the performance and the expectations of the investor's index fund replicating portfolio.

Whenever an index rebalances or makes changes stocks are routinely repriced at a substantial premium to market valuations multiples whereas discretionary deletions are routinely deep-discount value stock (Anott, Brightman and Kalesnik). (Beneish and Whaley) find that the returns to a stock that was added to or deleted from the S&P 500 from October 1989 through June 1994 did not adjust fully immediately after the announcement. They conjecture that this phenomenon will "disappear." However, (Blume and Edelen) found out the phenomenon diminishes instead of disappearing. (Swedroe) also studied top 500 stocks selected by market cap compared with S&P500 and the observation was that S&P500 had a larger tracking error. All these clearly show that there are changes, tracking errors and delays associated with index funds and its subindices which should be minimized as much as possible. (Blume and Edelen) put it clearly that an indexer that wishes to maintain tracking errors of the small magnitudes obtained must invest in ways that closely approximate an exact replication strategy. Specifically, such indexers must make most of their adjustments to changes in the index at or close to the closing price on the day of the change

Much as replication or sampling is concerned, an investor should not forget about the additional fees that come up during the process.

This study intends to minimize cost arising from fees by sampling a subset of subindices an index fund and the study actually tends to replicate the prices and/returns of the index by way

of modeling time series of a subset of correlated indices this is similar to method applied by (Dyer and Guest) whereby they manually read the PIS in funds' 497Ks. However, we intend to use readily available data that suit varieties of investors and desired trading period of investors. We import price data of subindices that majorly correlates with the index fund. We strongly believe that by analyzing time series of correlated funds can mitigate unexpected huge tracking errors that arise from the changes since stock expected to have changes can be dropped and the model stays nearly stable.

The study shall somewhat be similar to that of (Blume and Edelen), however our study focuses on studying subindices of the index fund unlike that of (Blume and Edelen) that studies other indices to replicate an index fund.

3 Study Materials and Methods

3.1 Scope of the Study

3.1.1 Time Scope:

The study shall be conducted on daily price data for 10 years. The first 6 years shall be used as in sample datasets for the time series modeling of the sub-indices whereas the remaining 4 years data shall be used as out of sample dataset for the time series modeling of the sub-indices. This shall be referred to as Phase 1 modeling.

The phase 1 out of sample predictions for sub-indices shall be recombined with index fund actual prices to form raw data for Phase 2 modeling. The first 3 years of phase 2 raw data shall constitute the in-sample data set whereas the last 1 year shall constitute the out of the sample data set.

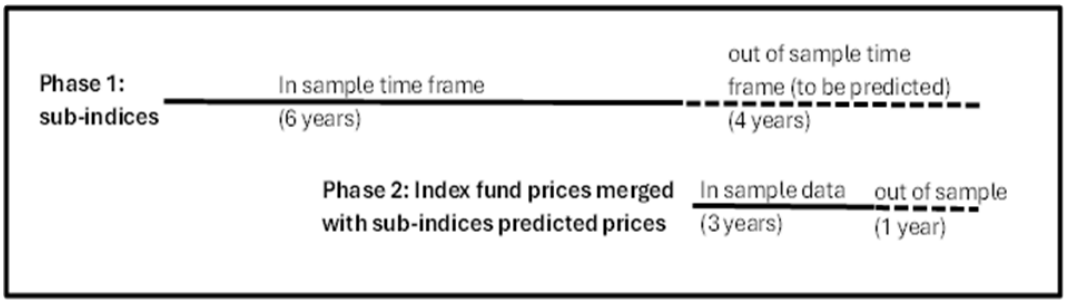


Figure 1: Time Scope Process Illustration

First of very all, we have extracted the list of all Dow Jones Industrial Average sub-indices with their full names, symbol (ticker) and year when they were added to the DJIA index, the table below elaborate on it:

#	Company	Symbol	Year (Added)
1	Amazon.com	AMZN	2024
2	Salesforce	CRM	2020
3	Amgen	AMGN	2020

4	Honeywell	HON	2020
5	Dow	DOW	2019
6	Apple	AAPL	2015
7	NIKE	NKE	2013
8	Visa	V	2013
9	Goldman Sachs	GS	2013
10	UnitedHealth Group	UNH	2012
11	Cisco Systems	CSCO	2009
12	The Travelers Companies	TRV	2009
13	Chevron	CVX	2008
14	Verizon	VZ	2004
15	Microsoft	MSFT	1999
16	The Home Depot	HD	1999
17	Intel	INTC	1999
18	Johnson & Johnson	JNJ	1997
19	Walmart	WMT	1997
20	The Walt Disney Company	DIS	1991
21	JPMorgan Chase	JPM	1991
22	Caterpillar	CAT	1991
23	The Coca-Cola Company	KO	1987
24	Boeing	BA	1987
25	McDonald's	MCD	1985
26	American Express	AXP	1982
27	Merck & Co.	MRK	1979
28	IBM	IBM	1979

29	3M	MMM	1976
30	Procter & Gamble	PG	1932

Table 1: Complete List of DJIA Sub-Indices

3.1.2 Content Scope:

The study shall involve Dow Jones Industrial Average sub-indices whose prices are highly correlated with the price of the index fund (Dow Jones Industrial Average) within the training set. The study shall consider prices for top 5 subindices which are most correlated with the index fund price. The study shall exclude subindices that have not spent 10 years by 31st December 2023 as a sub-index of Dow Jones Industrial Average.

After excluding the sub-indices which have not spent 10 complete years by 31st December 2023, we are left with 25 (out of 30) sub-indices highlighted in the following table:

#	Company	Symbol	Year (Added)
1	Company	Symbol	Year Added
2	NIKE	NKE	2013
3	Visa	V	2013
4	Goldman Sachs	GS	2013
5	UnitedHealth Group	UNH	2012
6	Cisco Systems	CSCO	2009
7	The Travelers Companies	TRV	2009
8	Chevron	CVX	2008
9	Verizon	VZ	2004
10	Microsoft	MSFT	1999
11	The Home Depot	HD	1999
12	Intel	INTC	1999
13	Johnson & Johnson	JNJ	1997
14	Walmart	WMT	1997
15	The Walt Disney Company	DIS	1991
16	JPMorgan Chase	JPM	1991

17	Caterpillar	CAT	1991
18	The Coca-Cola Company	KO	1987
19	Boeing	BA	1987
20	McDonald's	MCD	1985
21	American Express	AXP	1982
22	Merck & Co.	MRK	1979
23	IBM	IBM	1979
24	3M	MMM	1976
25	Procter & Gamble	PG	1932

Table 2: List of DJIA Sub-Indices which Spent 10 Years

3.2 Data Retrieval

The 30 companies that make up Dow Jones index fund as of July 2024 shall be extracted through web scraping with python panda's html reader (pandas.read_html, 2024). The web page that will be scrapped shall be for Investopedia and the content will be the Dow Jones companies posted on 7th July 2024 (Chen, 2024).

We have already extracted the list of DJIA sub-indices above (Table 1), where we have explicitly named all the 30 sub-indices with their full company names, symbols (tickers) and the year on which the sub-index was added to the main DJIA index.

The list will comprise of company name, symbol/ticker and the year the company was added to Dow Jones index fund. Only companies which were added within a period not less than 10 years will be used in sub-indices selection for model training. For details on this particular step, please look at Table 2 (above).

Adjusted closing prices for the subindices and for index fund (Dow Jones Industrial Average) shall be downloaded through yahoo finance module in python ("yfinance", 2024). As mentioned, we have retrieved the data for all the sub-indices along with the data for the main DJIA index starting from 1st January 2014 till 31st December 2023 (within the Source Code file).

3.3 Variables/Sub-Indices Selection

Pearson product-moment correlation coefficient (r) shall be used to determine the sub-indices which are most correlated with the index fund (Dow Jones Industrial Average). Pearson product-moment correlation coefficient (r) formula that shall be used to determine the correlation is shown below:

$$r = \frac{\Sigma(x_i - \bar{x})(y - \bar{y})}{\Sigma(x_i - \bar{x})^2 (y - \bar{y})^2}$$

Where x_i is the i^{th} independent variable (sub index), y is the dependent variable (Dow Jones Industrial Average), $\bar{x}$ and $\bar{y}$ are means for independent and dependent variables respectively (NLC, 2024). The study shall consider the top 5 most correlated variables as independent variables and the index fund price as the dependent variable.

The feedback from the instructor was to use PCA instead correlation analysis, which we have incorporated into the analysis and selected the top 5 PCA drivers from the list of sub-indices:

Sub-Index	UNH	V	MCD	JPM	MSFT
Correlation	0.766140	0.0.766050	0.764580	0.763454	0.762361

Table 3: Top 5 Sub-Indices based on PCA Explained Variance

3.4 Data Cleaning Process

The data shall be checked for missing values and in case there are any, the study shall consider dropping the missing rows provided they are less than 5% of the rows of the entire dataset when all variables are combined into a single data frame. In case the columns with null values are greater than or equals to 5%, the nulls will be replaced through filling them with the most preceding record. This is majorly aimed at reducing huge variations/swings within a short duration of time which are associated with replacing nulls with means.

Fortunately, there are no missing values looking at the Adjusted Price for all the 5 selected sub-indices based on their correlation coefficients.

3.5 Phase 1: Data Preprocessing and Time Series Analysis

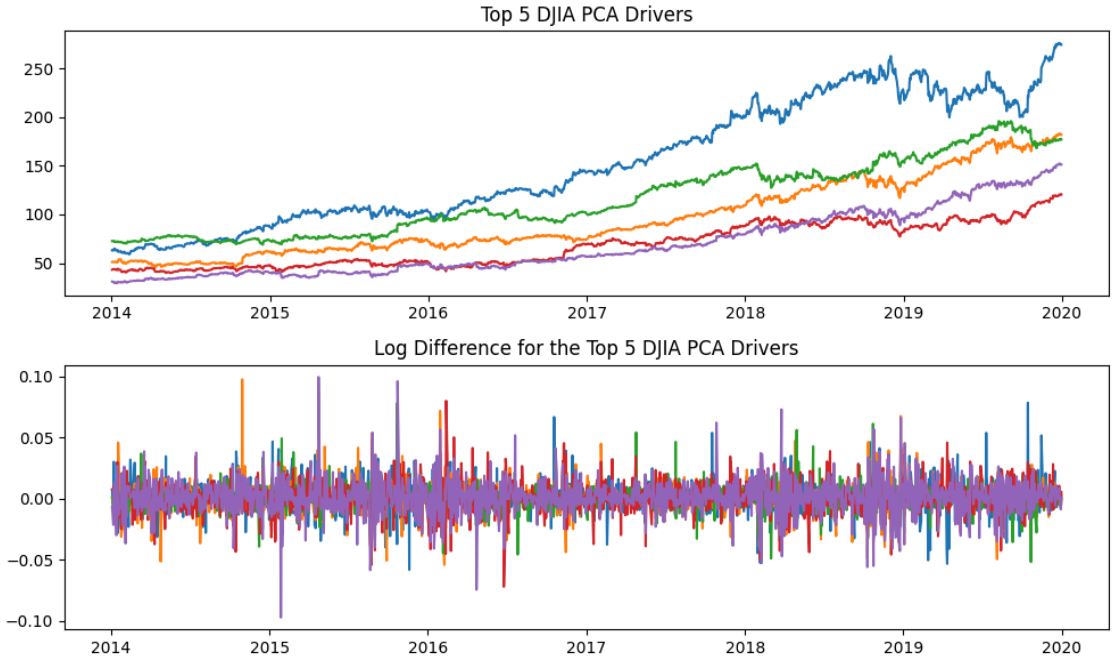
Each sub-index shall undergo individual time series analysis and prediction as follows:

3.5.1 Input Variables:

The sub-index daily prices shall be converted into a daily log return data. The log return shall be used majorly because it compresses large amounts of variations/jumps into a smaller range as per the advice by (Rawle, 2023). This is expected to smoothen data during a regime switch since the scope of this study does not incorporate regime switch in the time series modeling.

First, we have prepared the data for the selected sub-indices for phase 1 and then splitted that data into the training and testing sets. We will use the training data for training individual models and the testing data for out of sample testing purposes for evaluating the model performance. To reiterate, the training set starts from 1st January 2014 till the end of 2019 which is 6 complete years as mentioned above, and the testing data starts from 1st January 2020 till the end of 2023 (4 years).

Basic trends and the daily returns for the selected indicators of phase 1 data will look like following:



3.5.2 Data Training, Validation and Testing:

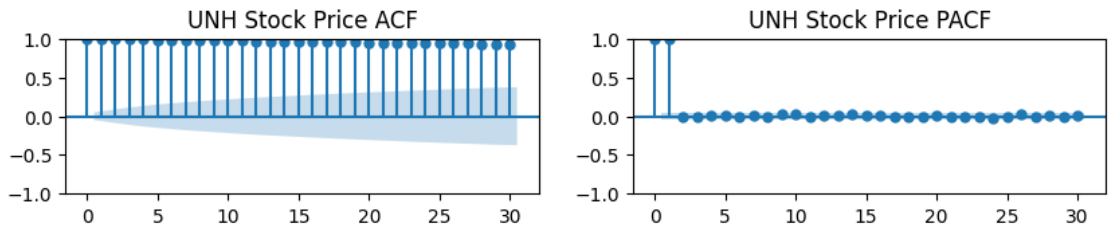
Several time series analysis techniques such as Auto-Regressive Integrated Moving Average (ARIMA), Generalized Autoregressive Conditional Heteroskedasticity (GARCH), Long Short-Term Memory (LSTM) among others shall be used and the model technique that output the lowest mean squared error shall be considered for sub-index modeling. Walk-Forward Optimization shall be used during modeling for better convergence of hyperparameters and improvement of model performance.

3.5.3 Modeling:

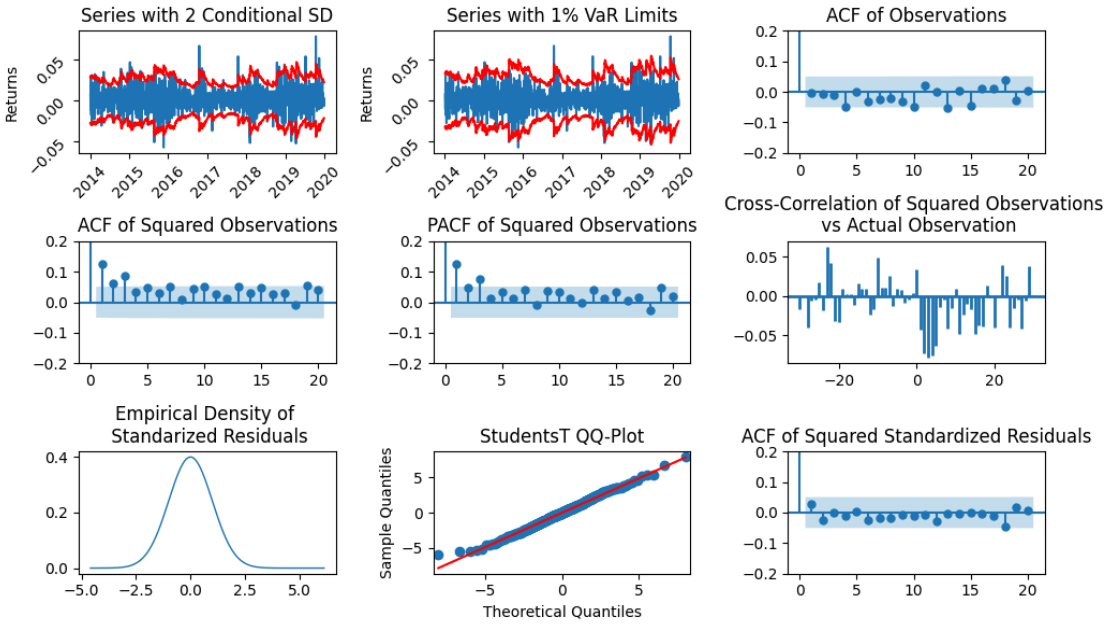
For modeling purposes, we have started with defining various functions where the first function looks at the trend, log differences, ACF, PACF plots…etc for the selected sub-index and the second function is looking at the GARCH diagnostics for that particular sub-index. The third function will model the data for the selected sub-index while the fourth one will provide the model predictions for the specified sub-index.

3.5.3.1 UNH Modeling and Results:

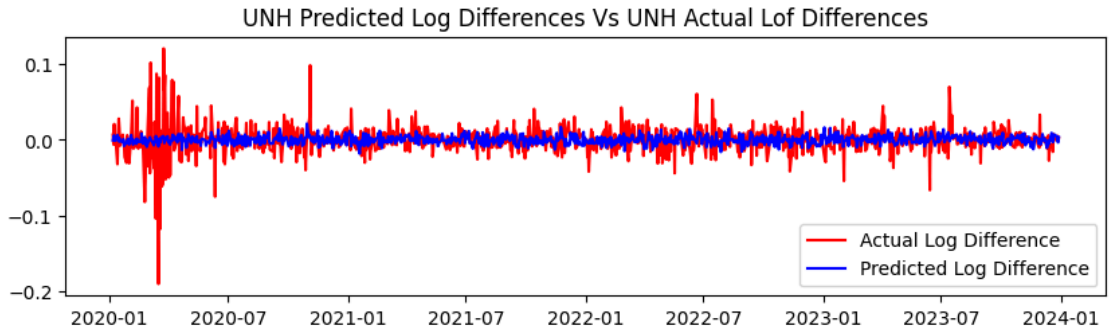
As mentioned above, the first functions looks at various trend and distribution related plots of the selected sub-index, in this case the UNH, out of which, I will only present the ACF and PACF which shows that there is an autocorrelation attached to the trend like following:



But the log differenced version of the trend is having no autocorrelation function to it, which we can notice on the source document. The next function is looking at various GARCH related functions for the particular sub-index like following:

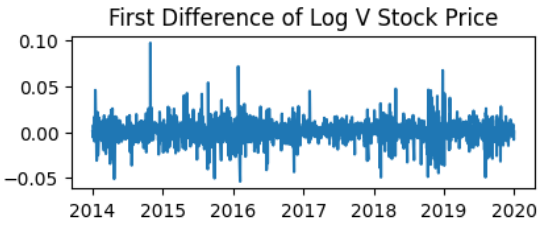


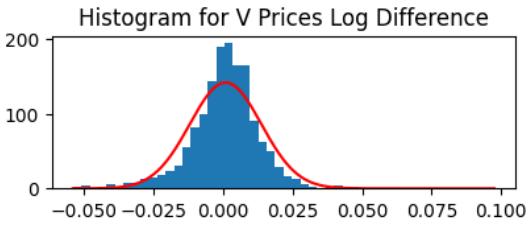
Finally, you can find the model summary on the source code document and the function will also provide the plot for comparing the actual and predicted log difference values for the selected sub-index.



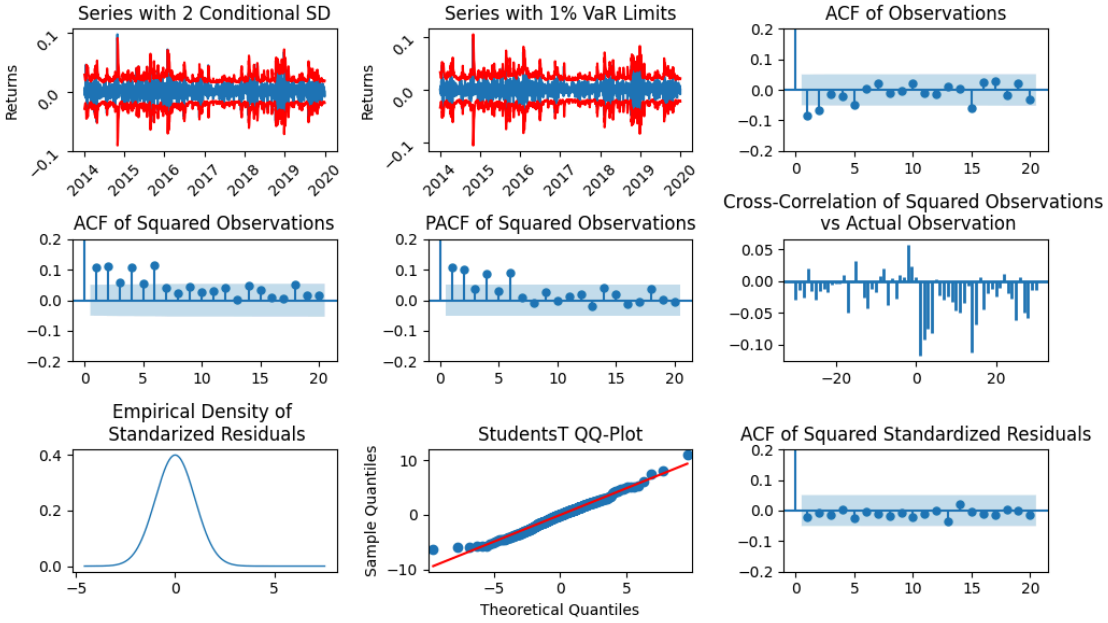
3.5.3.2 V Modeling and Results:

For V, we will present the first difference of log price and its histogram as an example from the very first function like following:

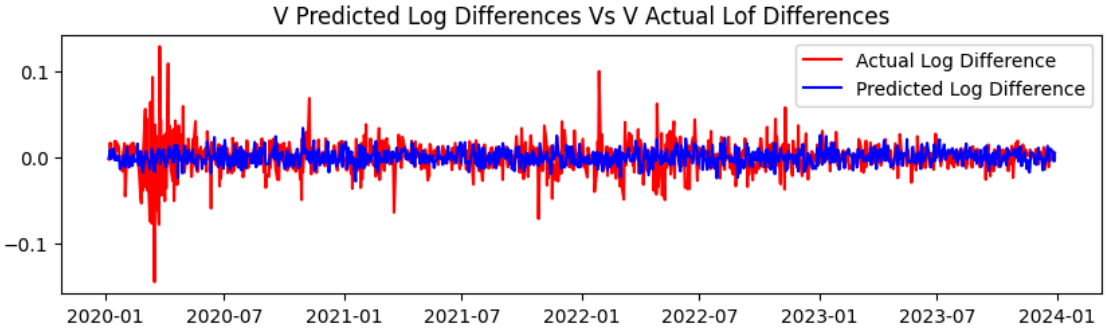




For the GARCH diagnostics, the V resulted in the following plots:

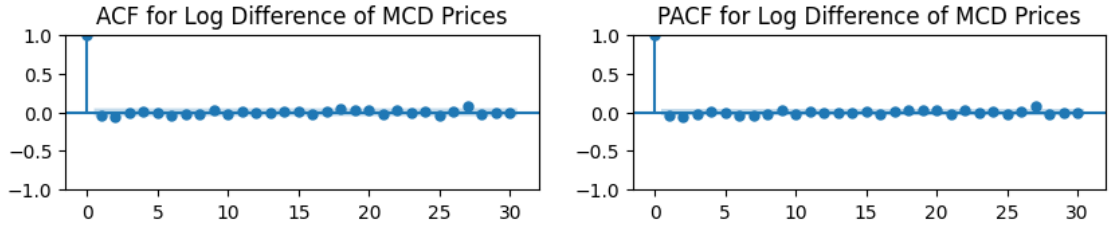


And the predicted vs actual comparison is as following:

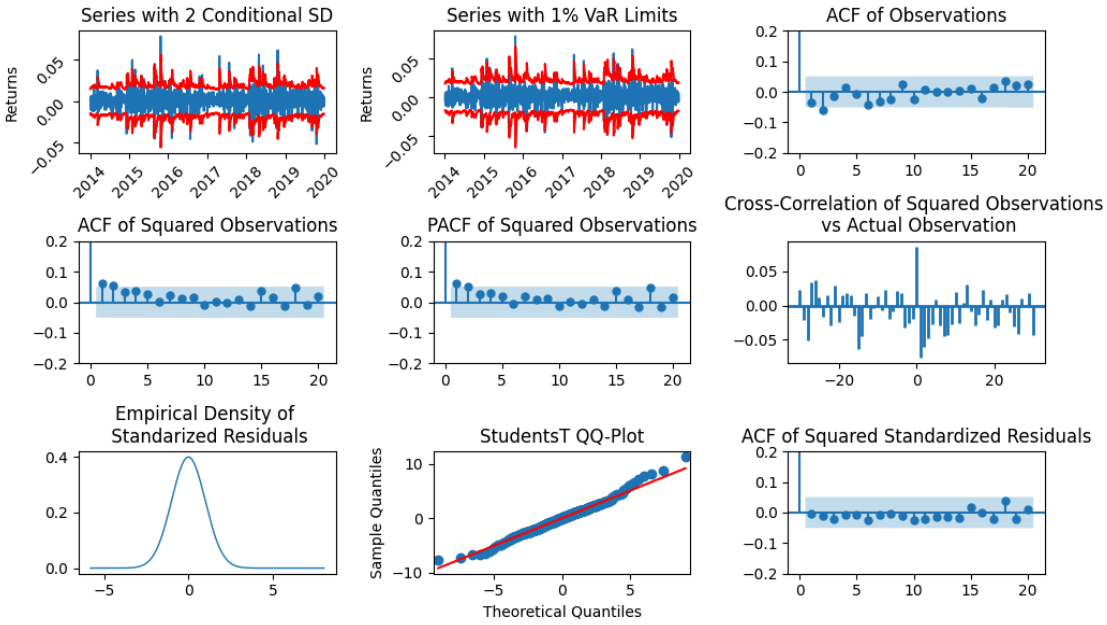


3.5.3.3 MCD Modeling and Results:

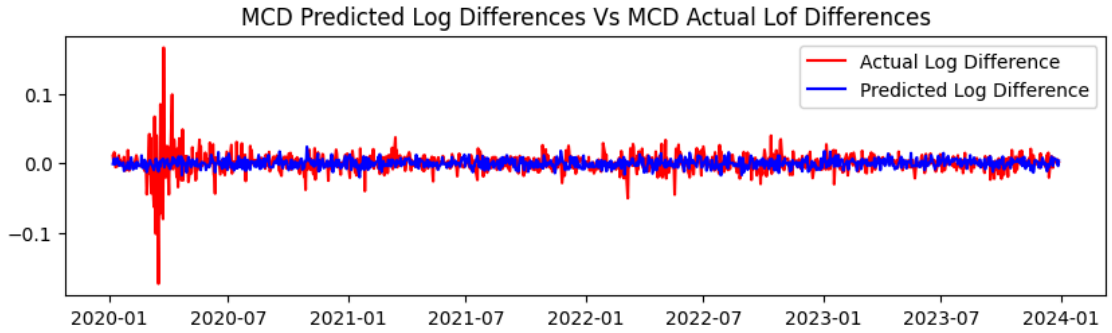
For MCD, we will only present the ACF and PACF of the log differences of its price, where it carries no autocorrelation to it after the first lag.



The diagnostics are as following:

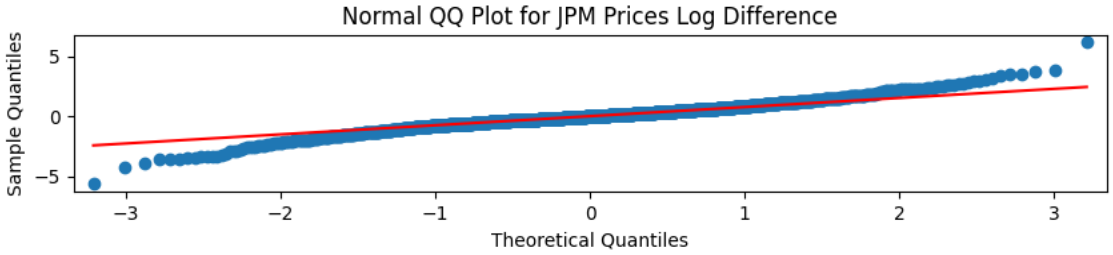


And the final plot which compares the actual and predicted log differenced values is as following:

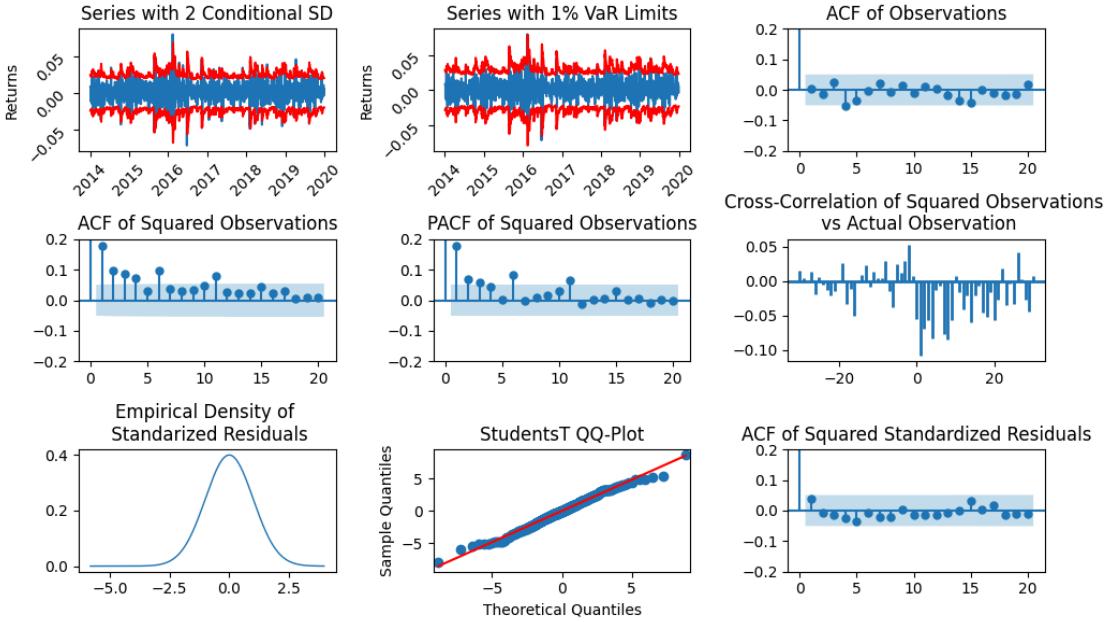


3.5.3.4 JPM Modeling and Results:

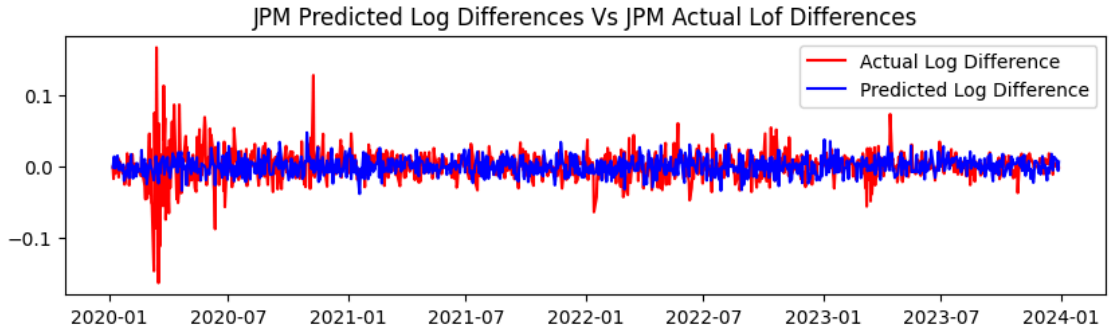
For JPM, we will only present the Normal QQ-Plot of the log differenced prices, the rest of the plots could be monitored through the source code document.



The GARCH diagnostics for the particular sub-index are resulting in following visuals:

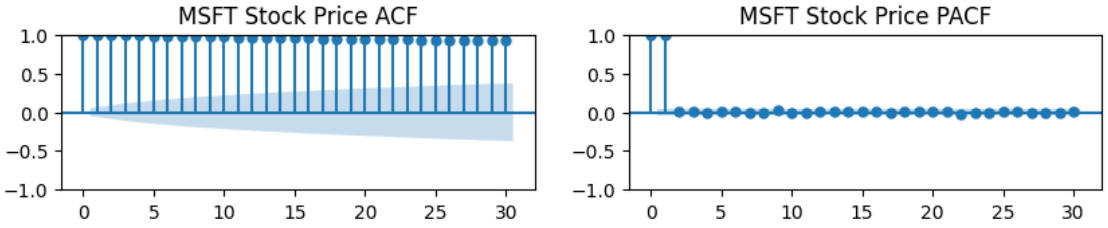


And the final plot for actual and predicted values comparison is as following:

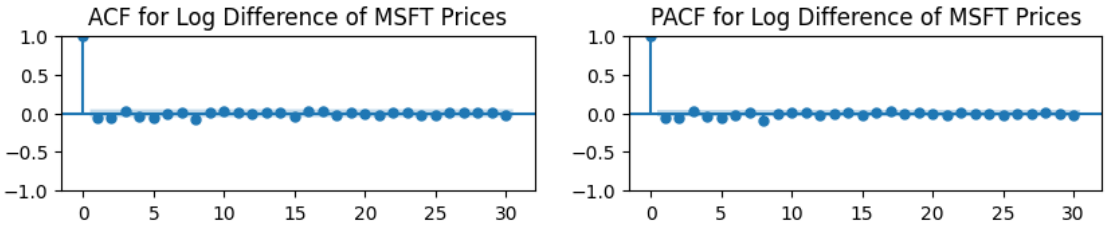


3.5.3.5 MSFT Modeling and Results:

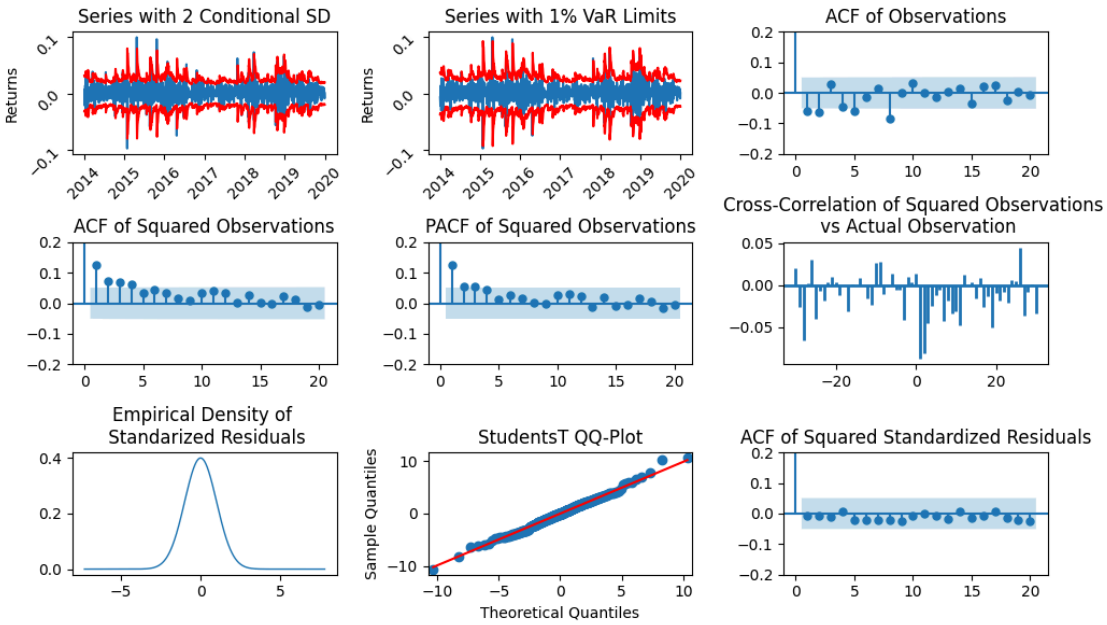
For MSFT, let's look at the ACF and PACF for price first:



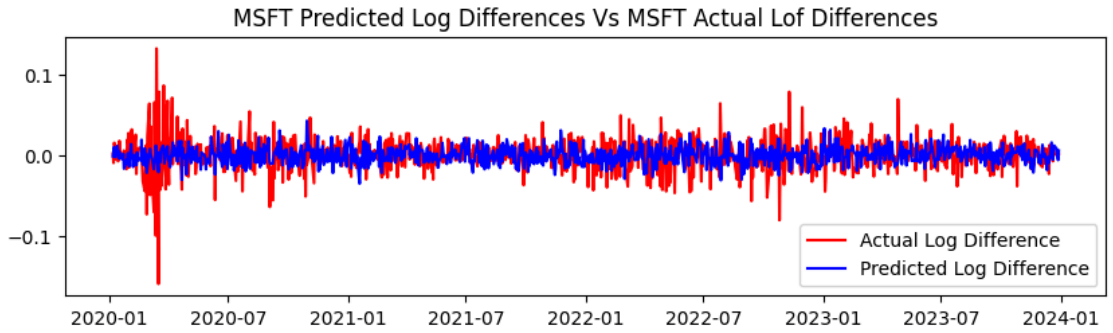
And then the same ACF and PACF for the log difference to compare it to the previous results:



After that, let's have a look at the GARCH diagnostics for the selected sub-index:



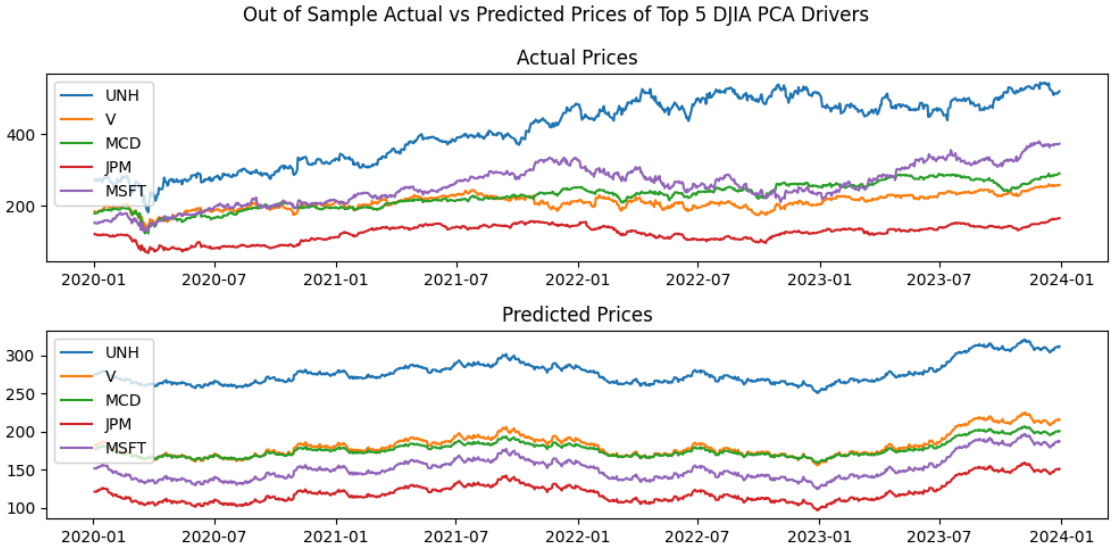
And finally, the comparison:



3.5.3 Output Data:

The predicted output of the log returns shall be converted back to prices by getting their exponential values. This output data shall be considered as input data for the next phase (index fund replicating model).

After going through the individual modeling process for each of the selected sub-indices, we have compiled back the predictions data from all individual models and compared it to the testing set that we kept aside while we were splitting the data for modeling and validation purposes. The actual test results vs the predicted ones will look like following:



We will use the same predicted values to replicate the overall DJIA index in the next phase.

3.6 Phase 2: Index Fund Replication Model Building

The prediction results from the previous phase and the index fund price data shall be combined into a single data frame.

3.6.1 Input Data:

The prediction results shall constitute the input independent variables, and the index fund actual price data shall constitute the dependent variable. Predicted prices shall be considered for the model training. Prices are believed to be a good predictor as observed from study of

Price Vs Return in Financial Forecasting with Machine Learning where the study found out that "that stock price is a more effective standalone input feature than return" in classification algorithms (Kamalov, Gurrib, & Rajab, 2021).

3.6.2 Data Preprocessing:

To balance the impact of all variables and to improve the performance of the algorithm, the independent variables shall be rescaled into a range of 0 to 1 (Khoong, 2023).

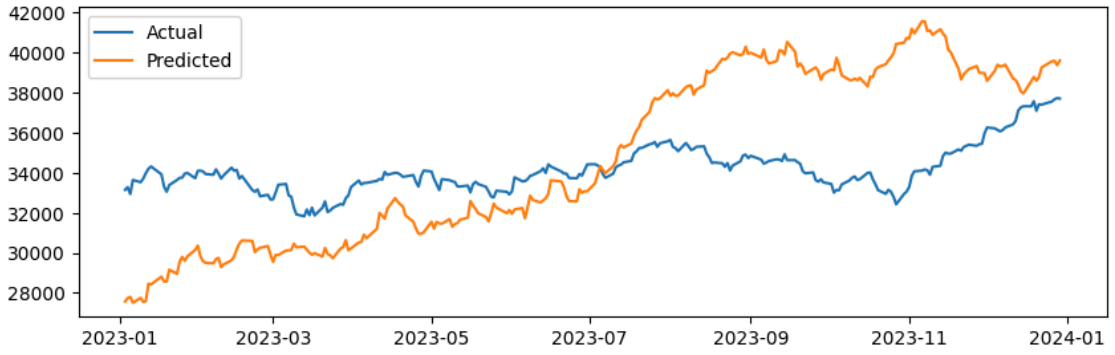
We have used the Min Max Scaler to rescale the data in the range of 0 to 1.

3.6.3 Model Training:

Regression analysis will be employed to predict the actual price of the index fund. Several modeling methods consisting of decision Trees, support vector machines, logistics regressions among others will be used to predict the index fund price and whichever output the lowest Mean Squared Error will be considered for the final model predictions. K-fold cross validation will be used during training to enhance model performance.

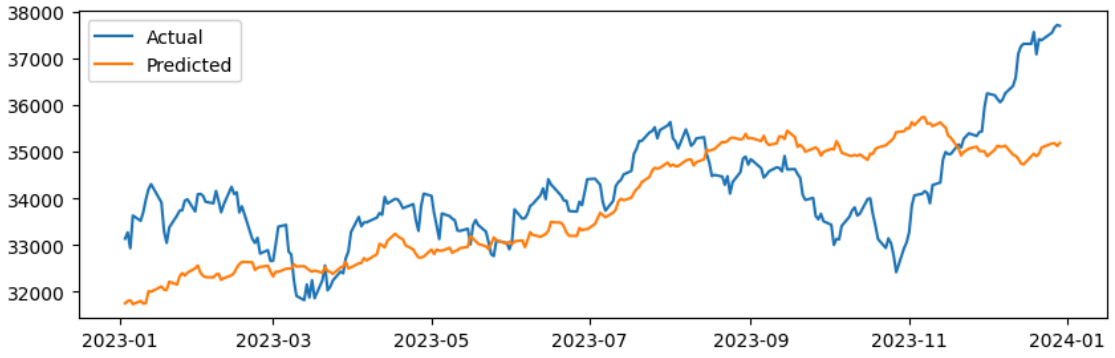
3.6.3.1 Linear Regression:

We have started modeling the data with linear regression where we have looked for the best parameter for the model first and used those parameter to model the data and compared the model predicted results with the actual one, but the results are no so satisfying:



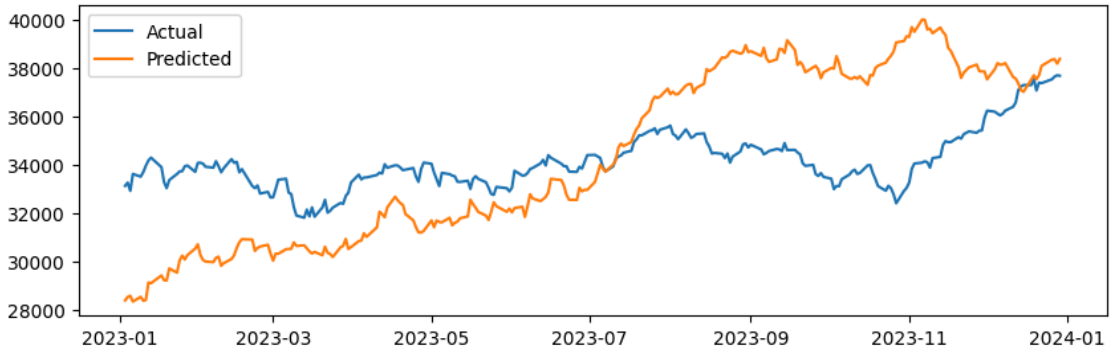
3.6.3.2 Ridge Regression:

After linear regression, we have looked through the best parameters for the ridge regression using grid search cross validation and modeled the data, then compared the model predicted results with actual ones, but unfortunately, the results were sort of similar (not satisfying) to that of the linear regression:



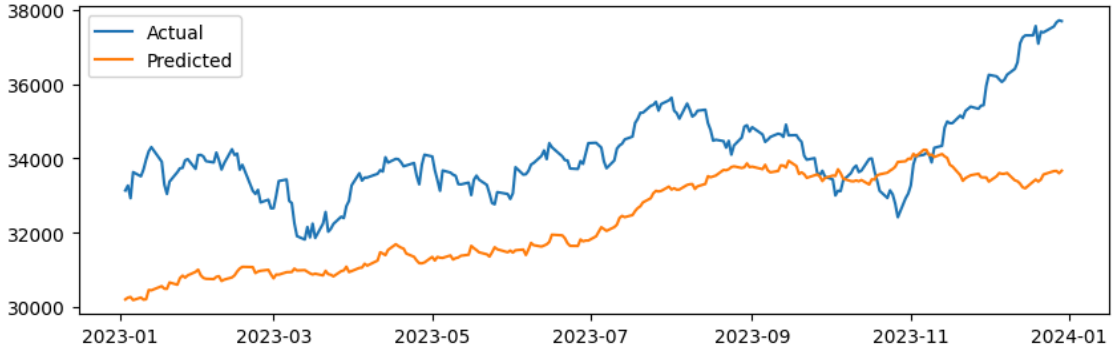
3.6.3.3 Lasso Regression:

This time, we have used grid search cross validation to search for best parameters for lasso regression and modeled the data, but when compared the model predicted results with the actual ones, the results were very similar to that of the previous models:



3.6.3.4 ElasticNet Regression:

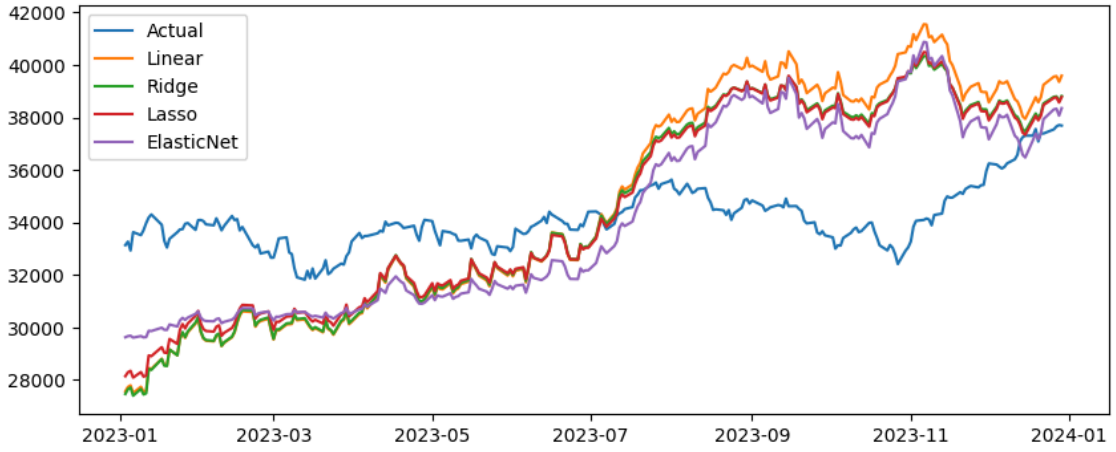
Finally, we have tried to used grid search cross validation to search for best parameter for ElasticNet Regression model and modeled the data, still the results were not that satisfying:



3.6.3.5 Adding Polynomial Features:

After going through multiple linear regression models, we have tried to add polynomial features to the data and reuse all the previous model with the revised data having 2 degrees polynomial features and compared the predicted results for all the models with the actual

data where all the models were behaving sort of very likely to each other but could give satisfying results:



In the final project, we will explore and use more advanced regression models like SVM, Random Forest, Extreme Gradient Boost…etc which can handle outliers very easily compared to the traditional linear models to get the required results.

The predicted index fund will constitute the index fund replica.

3.7 Modeling Process Flow

The prediction results from the previous phase and the index fund price data shall be:

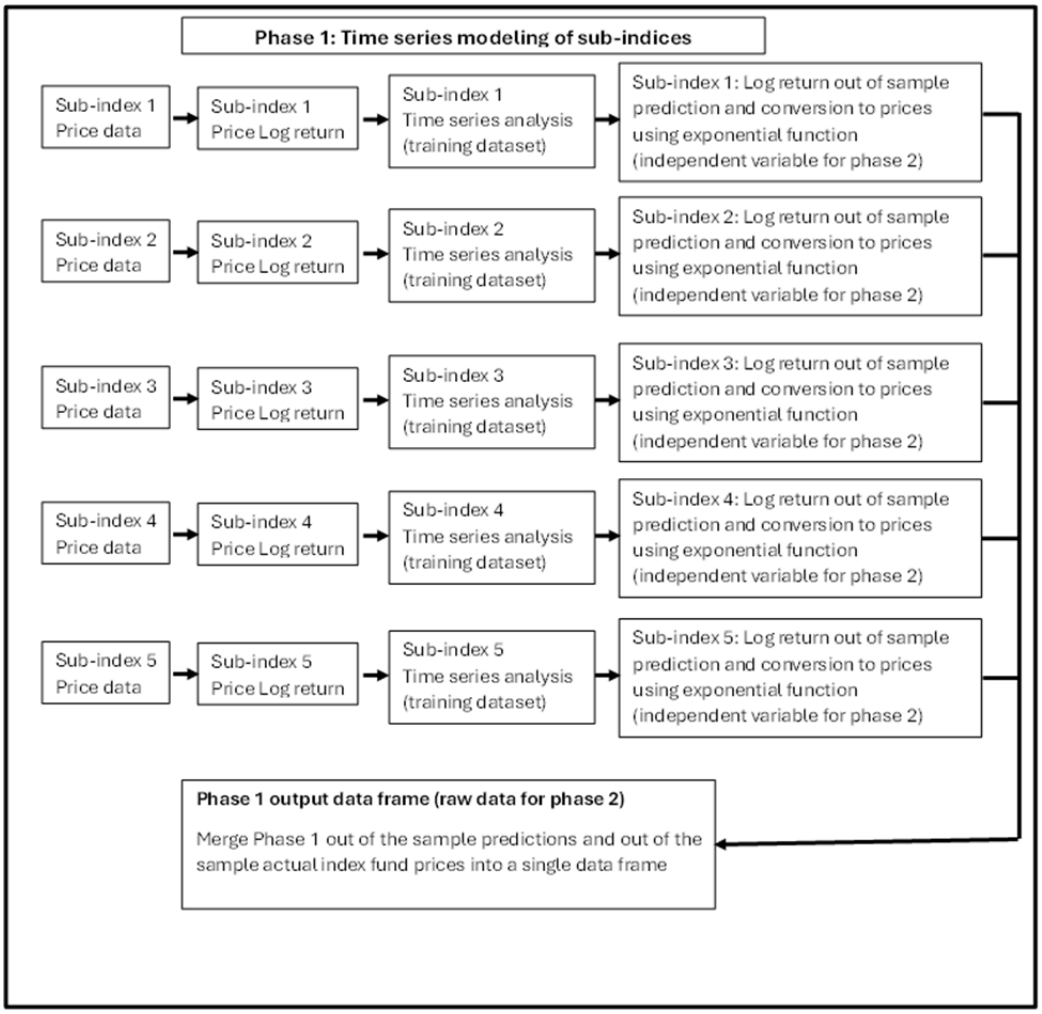


Figure 2: Time Series Analysis Process for the Sub-Indices

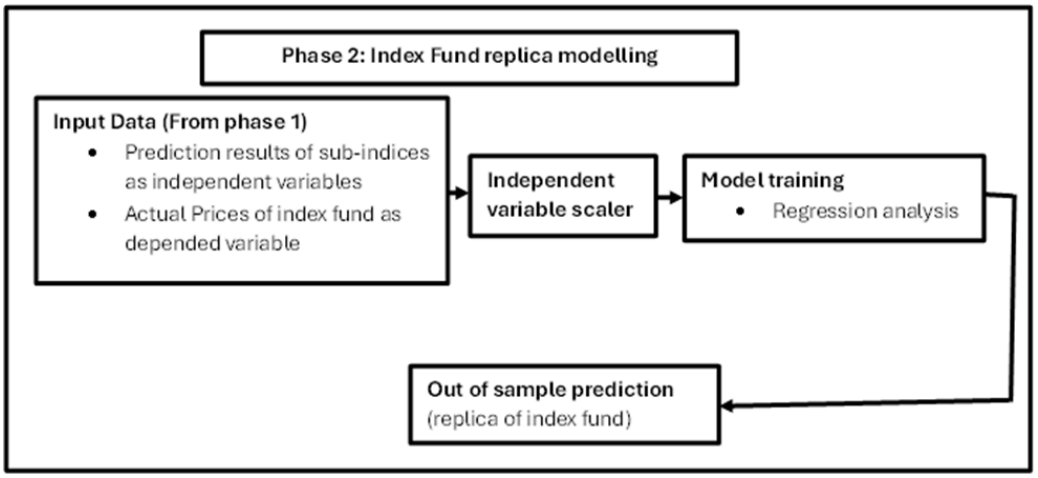


Figure 3: Modeling Process for the Index Fund Prices

Source Code Repository

https://github.com/mizewar/MScFE690_Capstone_Replication

Bibliography

[01]. Anott, Robert, et al. Earning Alpha by avoiding index rebalancing crowd. 2023.

[02]. Beneish, M. and R. Whaley. “The effect of changing the rules.” Journal of Finance 51 (1996).

[03]. Blume, Marshall E and Roger M. Edelen. “On Replicating the S&P 500 Index.” The Rodney L. White Center for Financial Research (2002).

[04]. Blume, Marshall E. and Roger M. Edelen. “On Replicating the S&P 500 Index.” The Rodney L. White Center for Financial Research (2002).

[05]. Blume, M. E., & Edelen, R. M. (2002). Replication Strategies. In On Replicating the S&P 500 Index

[06]. Charles SWAP. What Happens When the Stocks in an Index Change? 2022.

[07]. Chen, J. (2024, July 7). What Is the Dow 30? Companies In It, Significance. https://www.investopedia.com/. Retrieved from What Is the Dow 30? Companies In It, Significance.

[08]. Croome, Shauna. “The Hidden Differences Among Index Funds.” 12 May 2024. Investopedia. https://www.investopedia.com/articles/mutualfund/03/061103.asp. 30 August 2024.

[09]. Dyer, Travis and Nicholas Guest. “A Tale of Two Index Funds: Full Replication vs. Representative Sampling.” (2022).

[10]. Kamalov, F., Gurrib, I., & Rajab, K. (2021). Financial Forecasting with Machine Learning: Price Vs Return. Journal of Computer Science.

[11]. Khoong, W. H. (2023, January 21). Why Scaling Your Data Is Important. https://medium.com/codex/why-scaling-your-data-is-important-1aff95ca97a2.

[12]. NLC. (2024). Strength of Correlation. https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/strength-of-correlation.html.

[13]. pandas.read_html. (2024, 09 27). pandas. Retrieved from pandas: https://pandas.pydata.org/docs/reference/api/pandas.read_html.html

[14]. Rawle, A. (2023). Why do we use log-returns in financial time series modeling? Research gate. https://www.researchgate.net/post/Why_do_we_use_log-retuns_in_financial_time_series_modelling/6540f5d229405c3c6c079f6a/.

[15]. Swedroe, Larry. Improving Performance by Avoiding Negatives of Index Replication. 2023

[16]. yfinance. (2024, September 28). “yfinance”. Retrieved from “yfinance”: https://pypi.org/project/yfinance/