**GROUP WORK PROJECT #** 2
**GROUP NUMBER:** 3997

MScFE 610: FINANCIAL ECONOMETRICS


| FULL LEGAL NAME | LOCATION (COUNTRY) | EMAIL ADDRESS | MARK X FOR ANY NON-CONTRIBUTING MEMBER |
|---|---|---|---|
| Bharat Swami | India | bharatswami1299@gmail.com | |
| Ka Man Lui | United Kingdom | thomaslui.0924@gmail.com | |
| Daxin Niu | United States | daxinniu.work@gmail.com | |


| **Statement of integrity:** By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above). | |
|---|---|
| **Team member 1** | Bharat Swami |
| **Team member 2** | Ka Man Lui |
| **Team member 3** | Daxin Niu |

| Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.<br>**Note:** You may be required to provide proof of your outreach to non-contributing members upon request. |
|---|
| |

**Step 1**

**Multicollinearity**

Multicollinearity means that more than two explanatory variables are highly linearly correlated in the multiple regression model. In mathematical term, a set of variable is perfectly multicollinear if there are one or more than one linear relationships among some variables in the set.

$$\lambda_0 + \lambda_1 X_{1i} + \lambda_2 X_{2i} + \ldots + \lambda_n X_{ni} = 0$$

where $X_{ki}$ $is\ the\ i^{th}\ observation\ on\ the\ k^{th}\ explanatory\ variable\ and\ \lambda_k\ are\ constants$

**Unit Testing**

Unit testing is a method to check whether the given time series is stationary or non-stationary. The core principle to unit testing is to check for unit root in time series, if the unit root is present in the time series then it is a non-stationary time series, on the other hand if unit root is not present then we can say time series is stationary.

There is lots of unit test methods, following are the most popular one-

1. Augmented Dickey-Fuller Test (ADF)
2. KPSS Test
3. Ljung-Box Test

In all above tests we assume the null hypothesis to be that "Unit root is not present in the time series" i.e., Time Series is stationary.

**Feature Extraction**

Feature extraction is a common method used in math and financial domains. It captures the essentials of a dataset which allows us to narrow it down to a usable small dataset. In mathematical terms, suppose we are provided a dataset X where there are n different variables. Feature extraction allows us to extract a set of variables Y with m elements where $m \leq n$.

**Regime Change Models**

Regime Change Models refers to the model that categorize the data into different "state" or "regime". One of the examples is the autoregressive model which the mean of the process changes between regime according to the transition probabilities matrix in Hamilton's paper (Hamilton). The other example is the regime change model with variance switch in Kim, Nelson, and Startz's paper (Kim et al.)

## Step 2

## Multicollinearity

Description:

Multicollinearity is the situation when one explanatory variable in a regression model is highly correlated to the other explanatory variables.

Demonstration:

We want to study the relationship on movement between the Bitcoin and different equity stocks using multiple regression model. Price data from the Bitcoin, SPDR S&P 500 Trust ETF, Microsoft, Nvidia, Google, Amazon, Apple and Meta in the year 2022 were downloaded from the yahoo finance API. The data were transformed to the daily percentage price change. The daily percentage price change of the bitcoin was regressed against the price change of the chosen equity stocks using Ordinary Least Squares method. The result is shown below:

OLS Regression Results

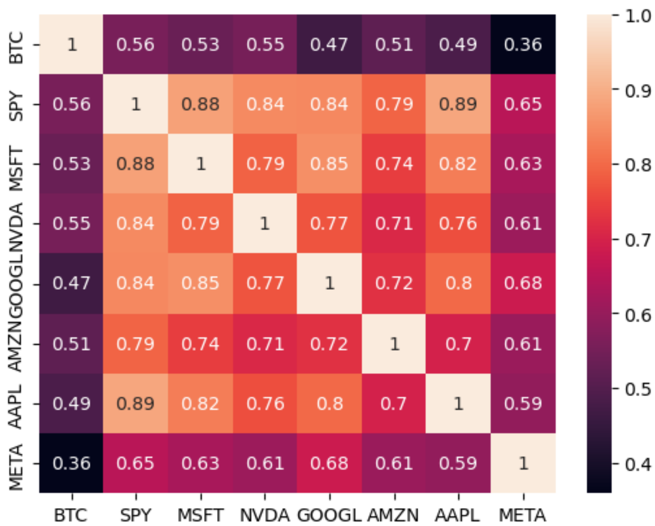
| | | | |
|---|---|---|---|
| **Dep. Variable:** | BTC | **R-squared:** | 0.355 |
| **Model:** | OLS | **Adj. R-squared:** | 0.336 |
| **Method:** | Least Squares | **F-statistic:** | 19.03 |
| **Date:** | Thu, 07 Sep 2023 | **Prob (F-statistic):** | 3.52e-20 |
| **Time:** | 23:47:51 | **Log-Likelihood:** | 522.56 |
| **No. Observations:** | 250 | **AIC:** | -1029. |
| **Df Residuals:** | 242 | **BIC:** | -1001. |
| **Df Model:** | 7 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | -0.0015 | 0.002 | -0.792 | 0.429 | -0.005 | 0.002 |
| **SPY** | 0.4998 | 0.386 | 1.294 | 0.197 | -0.261 | 1.261 |
| **MSFT** | 0.3286 | 0.205 | 1.606 | 0.110 | -0.075 | 0.732 |
| **NVDA** | 0.2563 | 0.093 | 2.746 | 0.006 | 0.072 | 0.440 |
| **GOOGL** | -0.2211 | 0.172 | -1.283 | 0.201 | -0.560 | 0.118 |
| **AMZN** | 0.2101 | 0.103 | 2.031 | 0.043 | 0.006 | 0.414 |
| **AAPL** | -0.0784 | 0.191 | -0.410 | 0.682 | -0.455 | 0.298 |
| **META** | -0.0414 | 0.067 | -0.617 | 0.538 | -0.173 | 0.091 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 26.104 | **Durbin-Watson:** | 1.911 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 125.656 |
| **Skew:** | 0.040 | **Prob(JB):** | 5.18e-28 |
| **Kurtosis:** | 6.472 | **Cond. No.** | 211. |

Diagram:

Correlation matrix between the variables




Variance Inflation Factor

| | VIF |
|---|---|
| **SPY** | 9.400407 |
| **MSFT** | 5.602537 |
| **NVDA** | 3.730029 |
| **GOOGL** | 4.757287 |
| **AMZN** | 2.871679 |
| **AAPL** | 4.972004 |
| **META** | 1.992641 |

Diagnosis:

To identify the existence of the problem, there are several ways to do so. First is to calculate the correlation matrix for the investigation of the multicollinearity. The correlation matrix of the dependent variable and independent variables is calculated for the dataset. From the correlation between the independent variables in matrix above, we can see that SPY is highly linearly correlated to other independent variables apart from META (i.e. around 0.8 - 0.9 for other independent variables). Regression model can be performed on SPY against other independent variables.

Out[5]:

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | SPY | **R-squared:** | 0.894 |
| **Model:** | OLS | **Adj. R-squared:** | 0.891 |
| **Method:** | Least Squares | **F-statistic:** | 340.2 |
| **Date:** | Mon, 11 Sep 2023 | **Prob (F-statistic):** | 3.51e-115 |
| **Time:** | 19:12:33 | **Log-Likelihood:** | 970.89 |
| **No. Observations:** | 250 | **AIC:** | -1928. |
| **Df Residuals:** | 243 | **BIC:** | -1903. |
| **Df Model:** | 6 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 0.0002 | 0.000 | 0.530 | 0.596 | -0.000 | 0.001 |
| **MSFT** | 0.1533 | 0.033 | 4.711 | 0.000 | 0.089 | 0.217 |
| **NVDA** | 0.0827 | 0.015 | 5.680 | 0.000 | 0.054 | 0.111 |
| **GOOGL** | 0.0382 | 0.029 | 1.339 | 0.182 | -0.018 | 0.094 |
| **AMZN** | 0.0785 | 0.016 | 4.782 | 0.000 | 0.046 | 0.111 |
| **AAPL** | 0.2438 | 0.028 | 8.820 | 0.000 | 0.189 | 0.298 |
| **META** | 0.0109 | 0.011 | 0.980 | 0.328 | -0.011 | 0.033 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 5.960 | **Durbin-Watson:** | 1.829 |
| **Prob(Omnibus):** | 0.051 | **Jarque-Bera (JB):** | 5.649 |
| **Skew:** | 0.340 | **Prob(JB):** | 0.0593 |
| **Kurtosis:** | 3.283 | **Cond. No.** | 116. |

$R^2$ and adjusted $R^2$ are high and it suggests that the info of the SPY daily return can be largely explained by other independent variables.

Another method is the variance inflation factor, it is used to test whether an independent variable has the issue of multicollinearity. The formula of VIF for the variable $X_j$ is as follows:

$$V_j = \frac{1}{1-R_j^2}$$

*where* $R^2_j$ *is the* $R^2$ *for the regression of* $X_j$ *on the other independent variables*

Higher the VIF for variable $X_j$, it shows that the multicollinearity of variable $X_j$ is more prominent.

From the above table, the VIF of the variable SPY is 9.4 and it shows that the variable shows a significant multicollinearity in the regression.

<u>Damage</u>:

The coefficients form the regression model may exhibit larger standard error and covariance if the independent variables are highly correlated. With the large standard error, the p-value of the coefficient estimate is smaller. With the smaller p-value, it is more difficult to identify the variables which should be included in the regression model as they are less likely to be statistically significant. The estimates are also less accurate due to the large standard errors. We will need to perform analysis on the variables with multicollinearity and see what actions need to be taken.

**Unit Testing**

Description: We are going to perform a unit test on our time series data to find whether the time series is stationary in nature or not. We are going to perform ADF and KPSS tests on time series with different lags and trends. Unit tests give Test-statistics and p-value with critical values at 1%, 5% and 10%.

For ADF: $H_0$ : Unit root is present in the time series; $H_1$: No unit root in time series

For KSPP: $H_0$: There is no trend in time series; $H_1$: There is a trend in time series

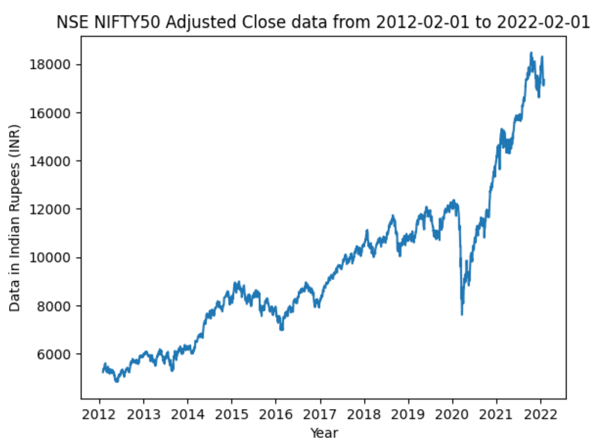
Demonstration: Again we are going to use NSE NIFTY50 index data as time series for our calculation (same used in GWP1 Kurtosis section). First we are going to download data using yfinance python module. We are again going to choose Adjusted daily close data for this index (labeled as “Adj Close”). Finally, we are going to perform ADF and KPSS unit tests on this data without any transformation. Below is the result from both unit tests.

| Sr. no. | Unit Test | | | Result** | | | | |
|---|---|---|---|---|---|---|---|---|
| | Test | Trend | Lags | Test-statistic | p-value | Critical values | | |
| | | | | | | 1% | 5% | 10% |
| 1 | ADF | c | 0 | 0.522 | 0.986 | -3.43 | -2.86 | -2.57 |
| 2 | KPSS | c | 0 | 203.429 | 0 | 0.74 | 0.46 | 0.35 |
| 3 | ADF | n | 0 | 2.411 | 0.0997 | -2.57 | -1.94 | -1.62 |
| 4 | ADF | ct | 0 | -1.557 | 0.809 | -3.96 | -3.41 | -3.13 |
| 5 | KPSS | ct | 0 | 15.223 | 0 | 0.22 | 0.15 | 0.12 |

Table : Adjusted Dickly-Fuller Test (ADF), Kwiatkowski-Phillips-Schmidt-Shin (KPSS)

c : Include a constant in Time series, n: No trend components, ct : Include a constant and linear time trend

Digram:



Diagnosis:
As we can see from the table in the description section, we can use ADF and KSPP functions from the arch python module. The output from these functions are test-statistics, p-value and critical values at 1%, 5% and 10% levels. The results can be interpreted as following cases-

- If the ADF test cannot reject the null hypothesis while KSPP can, then we can say the Time series is non-stationary.
- On the other hand, if ADF accepts the null hypothesis while KSPP cannot, then the Time series is stationary.

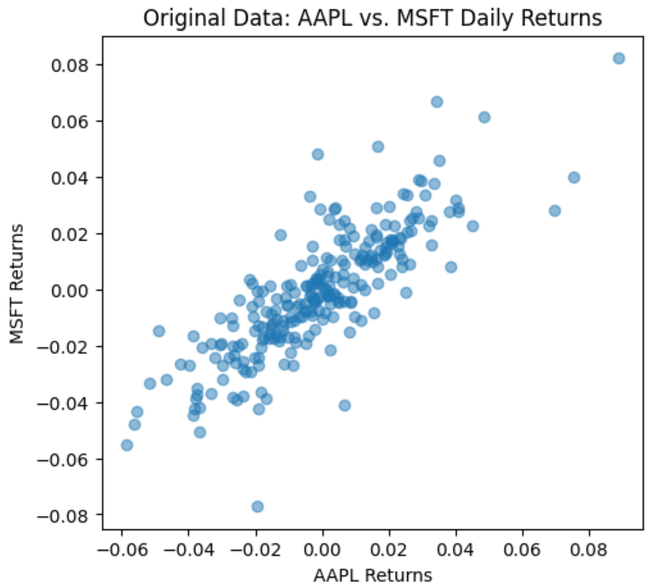
From the above p-values, we can say ADF accepts the null hypothesis and KSPP rejects it, which is Unit root present in Time series from ADF test and there is a trend in Time series from KSPP test. Which implies our time series data is non-stationary.
Damage:
If the time series is non-stationary, it will be difficult for us to model that time series for prediction of future values. Non-stationary time series are more complicated than stationary time series which use less computational power. It is advised to make time series stationary before proceeding with the modeling, and to make them stationary first we need to check whether they are non-stationary by using Unit Testing methods like ADF and KPSS.

**Feature Extraction**

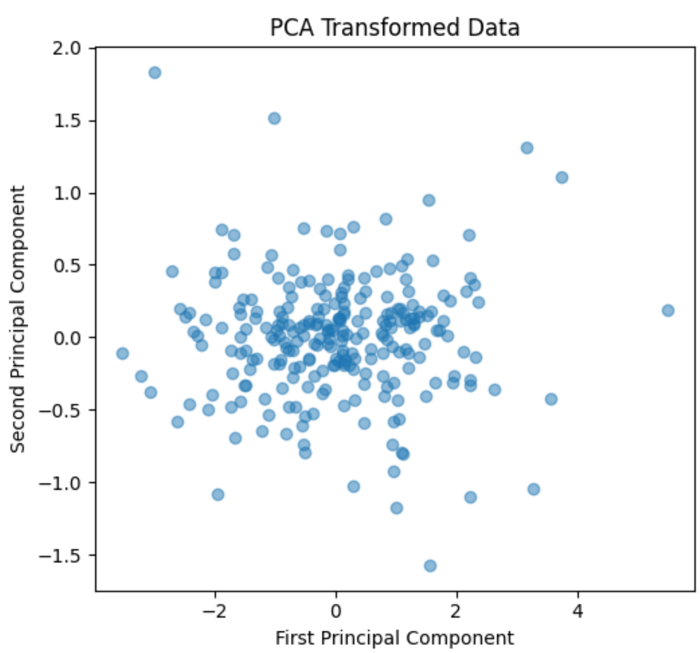
Description:
Feature extraction is quite necessary in many perspectives in finance. It allows us to better diversify portfolio by using only essential items from the data which can reduce the overall risk.

Demonstration:
To better illustrate the application of PCA, let us use a real world example. We have decided to use the data for APPL and MSFT over the period of 2022/01/01 to 2023/01/01 and take a look what we need to do if we want to build a portfolio using these two stocks. We first downloaded the data for these two stocks. Then we calculated the daily return of them. With the daily return of AAPL and MSFT, we plotted the data and realized that there is a high correlation between them. Then we decided to fit and transform the standard scaled return data using PCA. We were able to spot that the PC1 can explain over 90% of the variance. Lastly, we can use the PCA to get the eigenvalues and calculate the weight of the stocks in our portfolio.

Diagram:




The above illustration is the correlation mapping between AAPL and MSFT daily return from 2022/01/01 to 2023/01/01. We can see that the scatter plot is clustered together almost forming a line. This indicates that the two stock returns are highly correlated. In a situation like this, when we are building a portfolio using these two stocks, we need to consider doing feature extractions so that we can in some way reduce the overall correlation between assets.




Considering their high correlation, we have performed PCA on the data and reformatted them. The scatter plot of the PCA transformed data is shown above. We can see that the points are scattered around the 0 line on the y-axis. We have completed our PCA feature extraction here.

Diagnosis:

Following the PCA plots, we would want to check on the actual variance explained by using PCA. If we decide to print the explained variance ratio, we will get the following array.

```
1 pca.explained_variance_ratio_

array([0.91234541, 0.08765459])
```

We can see that over 90% of the variance and be explained by the first principle component. This is somewhat aligned with our original expectations. Considering both of the selected stocks have highly correlated daily returns, it is likely that one principle component will be able to cover a lot of the variance.

Although PCA looks great on the surface, there could still be issues that we need to take into consideration. As time goes on, the original principle component extracted might not fit the new data very well over time. Some highly correlated assets might at some point start to diverge. In such situations, it is necessary for us to keep track of our data quality and data correlations. It might be a good idea to perform feature extraction methods such as PCA when more data are ingested. Keeping the assumptions in check is usually very important in the financial world.

Damage:

The above section mentioned one of the common issues that could happen in the financial domain which is the failure of assumption. This is very common in the financial world. Take a reverse example from our demonstration. Suppose we built a portfolio with some highly uncorrelated assets A and B. Given that they have a lower correlation, we might have allocated quite a balanced amount of money in both assets for diversification purposes. Nonetheless, as time goes on, suppose there are some real-world events happening, and asset A starts to become highly correlated with asset B. If we didn't check their correlation time from time, we might still have the same distribution of these two assets in our portfolio. In such a situation, when they are highly correlated, we are no longer having the diversification benefit. In this case, if they are highly positively correlative, we could be taking big losses if one asset is going down since the over asset will likely decrease in value.

Considering the above situation, the damage could be quite big if we fail to keep track of our assumptions. Therefore, it is important that we keep track of the assumptions and adjust accordingly using our feature extraction methods.

**Regime Change Models**

Description: The Regime Change Models segment the data into different “states” or “regime”. The property of the data can change from one regime to another like the mean and variance.

Demonstration:
In this section we are again going with the NSE NIFTY50 index data from "2012-02-01" to "2022-02-01". And also we are going to use **Three-state Variance Switching** (Chad, 2013) method as Regime Change Model on our transformed data. We transformed our data, just like we did in Unit testing section, by using log transformation and then percentage return of adjusted close of index. As we established in the Unit testing “Direction” section that our transformed data is trend free which enables us to use Three-state Variance Switching Method. We are using the Markov Autoregression method (Josef Perktold, Skipper Seabold, Jonathan Taylor, statsmodels-developers ,2023 )to calculate the regimes, which are graphically presented in next section. The result summary from above method is shown below.

Markov Switching Model Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Adj Close | **No. Observations:** | 2448 |
| **Model:** | MarkovRegression | **Log Likelihood** | 8018.430 |
| **Date:** | Tue, 12 Sep 2023 | **AIC** | -16018.860 |
| **Time:** | 16:40:18 | **BIC** | -15966.633 |
| **Sample:** | 0<br>- 2448 | **HQIC** | -15999.879 |
| **Covariance Type:** | approx | | |

Regime 0 parameters

| | **coef** | **std err** | **z** | **P>\|z\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **sigma2** | 0.0001 | 6.52e-06 | 16.137 | 0.000 | 9.24e-05 | 0.000 |

Regime 1 parameters

| | **coef** | **std err** | **z** | **P>\|z\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **sigma2** | 3.891e-05 | 3.29e-06 | 11.829 | 0.000 | 3.25e-05 | 4.54e-05 |

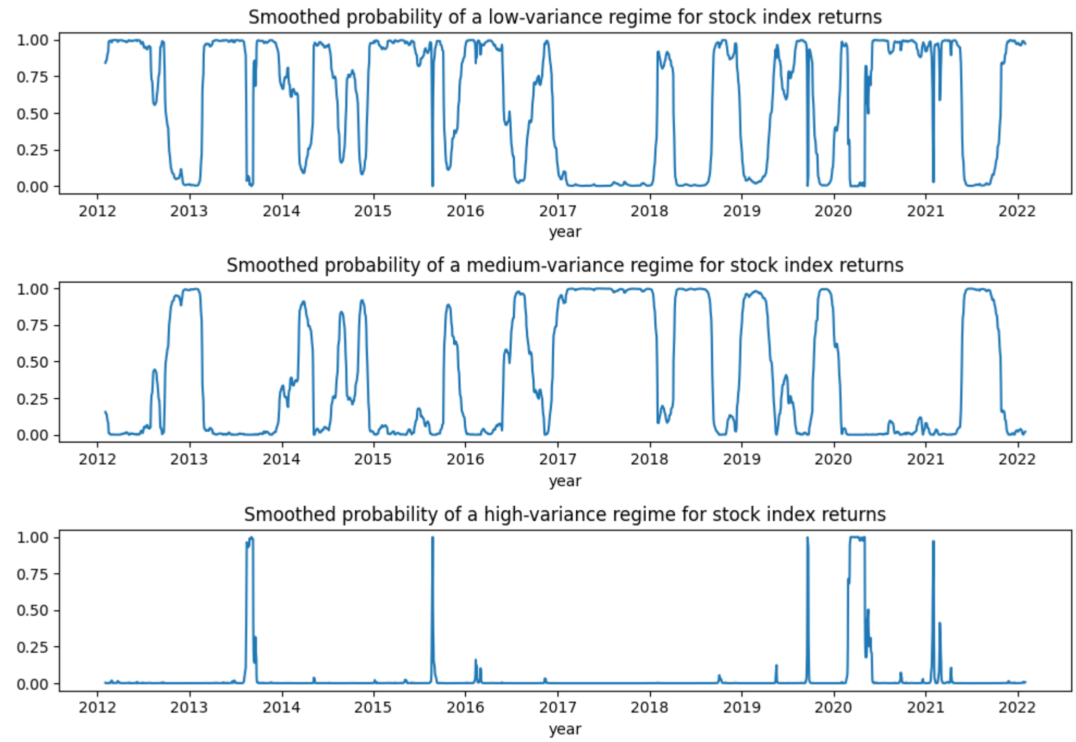
Regime 2 parameters

| | **coef** | **std err** | **z** | **P>\|z\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **sigma2** | 0.0012 | 0.000 | 4.781 | 0.000 | 0.001 | 0.002 |

Regime transition parameters

| | **coef** | **std err** | **z** | **P>\|z\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **p[0->0]** | 0.9832 | 0.005 | 203.312 | 0.000 | 0.974 | 0.993 |
| **p[1->0]** | 0.0172 | 0.000 | 90.065 | 0.000 | 0.017 | 0.018 |
| **p[2->0]** | 0.0963 | 0.043 | 2.252 | 0.024 | 0.012 | 0.180 |
| **p[0->1]** | 0.0113 | 0.004 | 2.821 | 0.005 | 0.003 | 0.019 |
| **p[1->1]** | 0.9828 | nan | nan | nan | nan | nan |
| **p[2->1]** | 3.481e-11 | nan | nan | nan | nan | nan |

Diagram:




From above graph we can see our data have some regimes divided into them. They can be helpful for us when we are modeling a model for stock index.

Diagnosis:

Regime change models are some of the strongest models. They can adapt to more macro movements and more granular levels of market movement. Nonetheless, there still exist some issues that could happen with regime change models. Regime selection could be one of the potential causes that could lead to some issues. With the wrong regimes, the model might not be capturing the market movement correctly. To identify such a situation, it might be best to test models with different regimes and observe model performances on the past data. With some backtests it could be easier to spot whether regimes are selected correctly.

Damage:

The damage of choosing the wrong regime could be devastating. Suppose we have selected a regime change model that helps track high volatility. During a recession or expansion economic environment, high volatility could be observed. But if we are using this model to track recession or expansion, it will not work correctly. On the surface, it might seem to have the correct prediction, nonetheless, it can completely get the wrong direction if we use it to solely predict either recession or expansion. We could see volatility go up and think it is an expansion, but it is very possible that it ended up turning out to be a recession. In a situation like this, we will be losing a great amount of money if we are betting in the wrong direction.

**Step 3**

**Multicollinearity**

Directions:

If we identify there's a variable with significant collinearity, we can run the regression model without that variable and see the impact. Since SPY is identified as the variable with significant collinearity in the dataset, SPY is removed and regression model has been run again.

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | BTC | **R-squared:** | 0.351 |
| **Model:** | OLS | **Adj. R-squared:** | 0.335 |
| **Method:** | Least Squares | **F-statistic:** | 21.86 |
| **Date:** | Mon, 11 Sep 2023 | **Prob (F-statistic):** | 1.60e-20 |
| **Time:** | 21:27:51 | **Log-Likelihood:** | 521.70 |
| **No. Observations:** | 250 | **AIC:** | -1029. |
| **Df Residuals:** | 243 | **BIC:** | -1005. |
| **Df Model:** | 6 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | -0.0014 | 0.002 | -0.747 | 0.456 | -0.005 | 0.002 |
| **MSFT** | 0.4052 | 0.196 | 2.066 | 0.040 | 0.019 | 0.792 |
| **NVDA** | 0.2977 | 0.088 | 3.389 | 0.001 | 0.125 | 0.471 |
| **GOOGL** | -0.2020 | 0.172 | -1.175 | 0.241 | -0.541 | 0.137 |
| **AMZN** | 0.2493 | 0.099 | 2.518 | 0.012 | 0.054 | 0.444 |
| **AAPL** | 0.0435 | 0.167 | 0.261 | 0.794 | -0.285 | 0.372 |
| **META** | -0.0359 | 0.067 | -0.536 | 0.593 | -0.168 | 0.096 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 25.687 | **Durbin-Watson:** | 1.910 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 121.763 |
| **Skew:** | -0.019 | **Prob(JB):** | 3.63e-27 |
| **Kurtosis:** | 6.419 | **Cond. No.** | 116. |

Compared to the original regression model, the p-values of the other variables increase by dropping the variable SPY and makes the coefficient estimate more statistically significant. The $R^2$ does not change a lot compared to the previous model.

The second method is to use principal component analysis to pick the variable with highest absolute value of the loading for each principal component as the proxy.

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|---|
| **Proportion of variance** | 0.785664 | 0.070136 | 0.047288 | 0.035582 | 0.029298 | 0.020275 | 0.011758 |
| **Cumulative proportion of variance** | 0.785664 | 0.855800 | 0.903088 | 0.938670 | 0.967967 | 0.988242 | 1.000000 |

The first 5 principal components were chosen for the proxy variable selection as they explain over 96% of the variance and the regression model was run using the variables selected and the result is as below:

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | BTC | **R-squared:** | 0.347 |
| **Model:** | OLS | **Adj. R-squared:** | 0.333 |
| **Method:** | Least Squares | **F-statistic:** | 25.89 |
| **Date:** | Mon, 11 Sep 2023 | **Prob (F-statistic):** | 6.04e-21 |
| **Time:** | 22:04:06 | **Log-Likelihood:** | 520.94 |
| **No. Observations:** | 250 | **AIC:** | -1030. |
| **Df Residuals:** | 244 | **BIC:** | -1009. |
| **Df Model:** | 5 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | -0.0014 | 0.002 | -0.741 | 0.460 | -0.005 | 0.002 |
| **SPY** | 0.6199 | 0.362 | 1.711 | 0.088 | -0.094 | 1.333 |
| **META** | -0.0565 | 0.065 | -0.875 | 0.382 | -0.184 | 0.071 |
| **AMZN** | 0.2167 | 0.103 | 2.113 | 0.036 | 0.015 | 0.419 |
| **NVDA** | 0.2583 | 0.092 | 2.808 | 0.005 | 0.077 | 0.439 |
| **AAPL** | -0.0664 | 0.186 | -0.358 | 0.721 | -0.432 | 0.299 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 26.157 | **Durbin-Watson:** | 1.931 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 124.901 |
| **Skew:** | 0.069 | **Prob(JB):** | 7.55e-28 |
| **Kurtosis:** | 6.460 | **Cond. No.** | 201. |

Compared to the original regression model, the p-values of the other variables increase and makes the coefficient estimate more statistically significant apart from AAPL. The $R^2$ does not change a lot compared to the previous model.

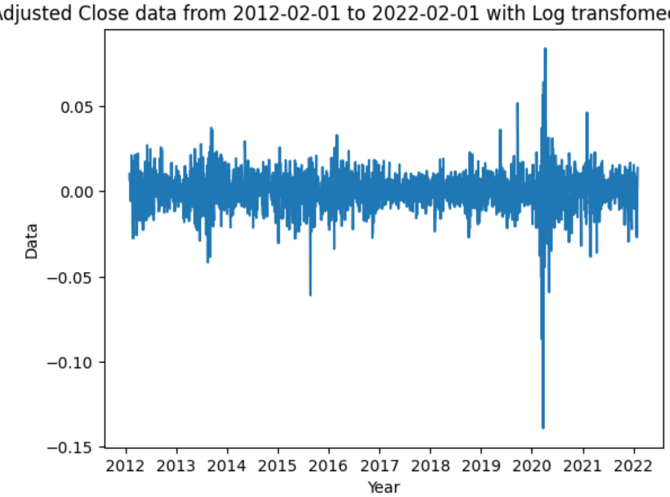
These two methods have tried to remediate the identification of the useful variables in the regression model when the multicollinearity exists.

## Unit Testing

Direction:

To solve the problem of non-stationary time series we can use different types of transformation by making them stationary. We can use lag difference, log transformation, square root transformation, etc. In our case we are going to use first log transformation then percentage difference of each data with previous data point. Below is the graphical representation of our transformed data.

NSE NIFTY50 Adjusted Close data from 2012-02-01 to 2022-02-01 with Log transfomed then percentage return



Now, as we can see we remove the upward trend from our data. Now, we perform the Unit tests again to see if this transformation changes anything.

| Sr. no. | Unit Test | | | Result** | | | | |
|---|---|---|---|---|---|---|---|---|
| | Test | Trend | Lags | Test-statistic | p-value | Critical values | | |
| | | | | | | 1% | 5% | 10% |
| 1 | ADF | c | 0 | -49.330 | 0 | -3.43 | -2.86 | -2.57 |
| 2 | KPSS | c | 0 | 0.048 | 0.889 | 0.74 | 0.46 | 0.35 |
| 3 | ADF | n | 0 | -49.240 | 0 | -2.57 | -1.94 | -1.62 |
| 4 | ADF | ct | 0 | -49.322 | 0 | -3.96 | -3.41 | -3.13 |
| 5 | KPSS | ct | 0 | 0.040 | 0.719 | 0.22 | 0.15 | 0.12 |

**Table**

As we can see from new test results we now reject the null hypothesis in ADF method and accept the null hypothesis in KSPP method. So, for this data this transformation makes it stationary.

*** please refer to attached python code*

**Feature Extraction**

Direction:
To solve the issue of change in assumption for feature extraction, some of the best directions are to keep track of the underlying data and check on assumption. To make sure our feature extraction is working correctly, we can apply several methods to make sure things are working correctly. Other than using PCA, we could also be using methods like factor analysis. By combining multiple methods together, we will be able to receive results with a bit higher confidence. Meanwhile, we should also keep track of our data and keep checking whether our assumption holds true. The reason why checking assumptions is important is that with new data, things can change quite drastically over time. If we use the data from 01/01/2022 to 01/01/2023, we have the following PCA explained variance ratio.

```
1 pca.explained_variance_ratio_
array([0.91234541, 0.08765459])
```

```
1 pca.explained_variance_ratio_
array([0.83998135, 0.16001865])
```

Nonetheless, if we shift the time from 01/01/2021 to 01/01/2022, we will get the second graph.
The results for PC1 changed quite drastically. Therefore, it is important that once in a while, we should double-check our assumptions.

**Regime Change Models**

Direction:
In order to see whether our model is working correctly, from the regime perspective, it might be better to keep tracking the model performance using different regimes. With multiple regime change models running at the same time, it could be a lot easier for us to track down whether our regime is working correctly.

**Multicollinearity and Feature Extraction**

Damage:

For multcollinearity, it is mainly about the imprecise coefficient estimate due to the large standard errors. The large standard errors make it difficult to decide which variables need to be included in the model as the estimates are less likely statistical significant.

For feature extraction, it aims to identify and extract the variables that are useful in the regression model and in someway help to diversify in many cases. It is quite similar to multicollinearity in someway that one would complement the other in some situations. Feature extraction can be seen as a method to reduce multicollinearity in many situations. The damage of these two terms are sometime offsetting each other which is quite interesting.

From the above, we can conclude that the interaction of the independent variables in the model can be the challenges in these two topics.

Directions:

From the analysis on the multicollinearity and feature extraction above, we can see that PCA can be one of the direction to remediate the issues.

For multicollinearity, we can choose the variable with highest absolute value of the loading for each principal component as the proxy We can select the first N principal components if the first N principal components explain over a certain percentage of variance. It can used as one of the way to perform analysis on the significance of the variables in the model and select the variables.

For feature extraction, we can perform PCA and identify the feature of the data. The feature of the data can be captured by the first N principal components as the top principal components captures the most of the variance. Some hidden interaction between the variables can be also discovered using PCA.

Reference

1. Hamilton, J. D. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57: 357–384
2. Kim, C.-J., C. R. Nelson, and R. Startz. 1998. Testing for mean reversion in heteroskedastic data based on Gibbs-sampling-augmented randomization. Journal of Empirical Finance 5: 131–154
3. Fulton, Chad (2013), http://www.chadfulton.com/topics/mar_kim_nelson_startz.html
4. Josef Perktold, Skipper Seabold, Jonathan Taylor, statsmodels-developers (2023) . statsmodels 0.14.0 .. https://www.statsmodels.org/stable/examples/notebooks/generated/markov_autoregression.html