**Statistical and regression modeling**

**for Nasdaq historic data between 2016 to 2020 in R**

Pengfei Ma

Boston University

CS 555

Prof. Heather

Feb.24th 2020

**Introduction**

The U.S. stock market was experiencing an unprecedented phenomenon, several fuses happened in 2020 but the price of three stock markets created another peak of history. In order to display the huge change of the U.S stock market, this project would choose the historic data for the past five years from 2016 to 2020 of Nasdaq stock market which was downloaded from Yahoo Finance. Statistical modeling and analysis will be applied to this data set.

**Research scenario and question**

Downloaded data set from *Yahoo Finance*. Named the file *Nasdaq.csv* and saved it to the local path. By running *compute\_stocks\_weekly\_return\_volatility.py*, generated two new files named *Nasdaq\_weekly\_return\_volatility\_detailed.csv* and *Nasdaq\_weekly\_return\_volatility.csv.* These three files will be the only files used as the data set in this project. Coding files *compute\_stocks\_weekly\_return\_volatility.py* will be only used to generate two new data sets, rather than data analytics. File *Data analytics of Nasdaq.R* will be the coding file that contains all codes of data analytics of this project.

Rstudio will be used as the only IDE for analytical works and programming language R will be the only programming language used for analysis. Subline Text will be used as IDE for python file to generate the data sets.

There are two questions came up before the project start:

1. What is the relationship between *Adj Close* and *Volume?* How strength of this relationship.
2. Are Adj Close and Return when considered together significant predictors of Volume?

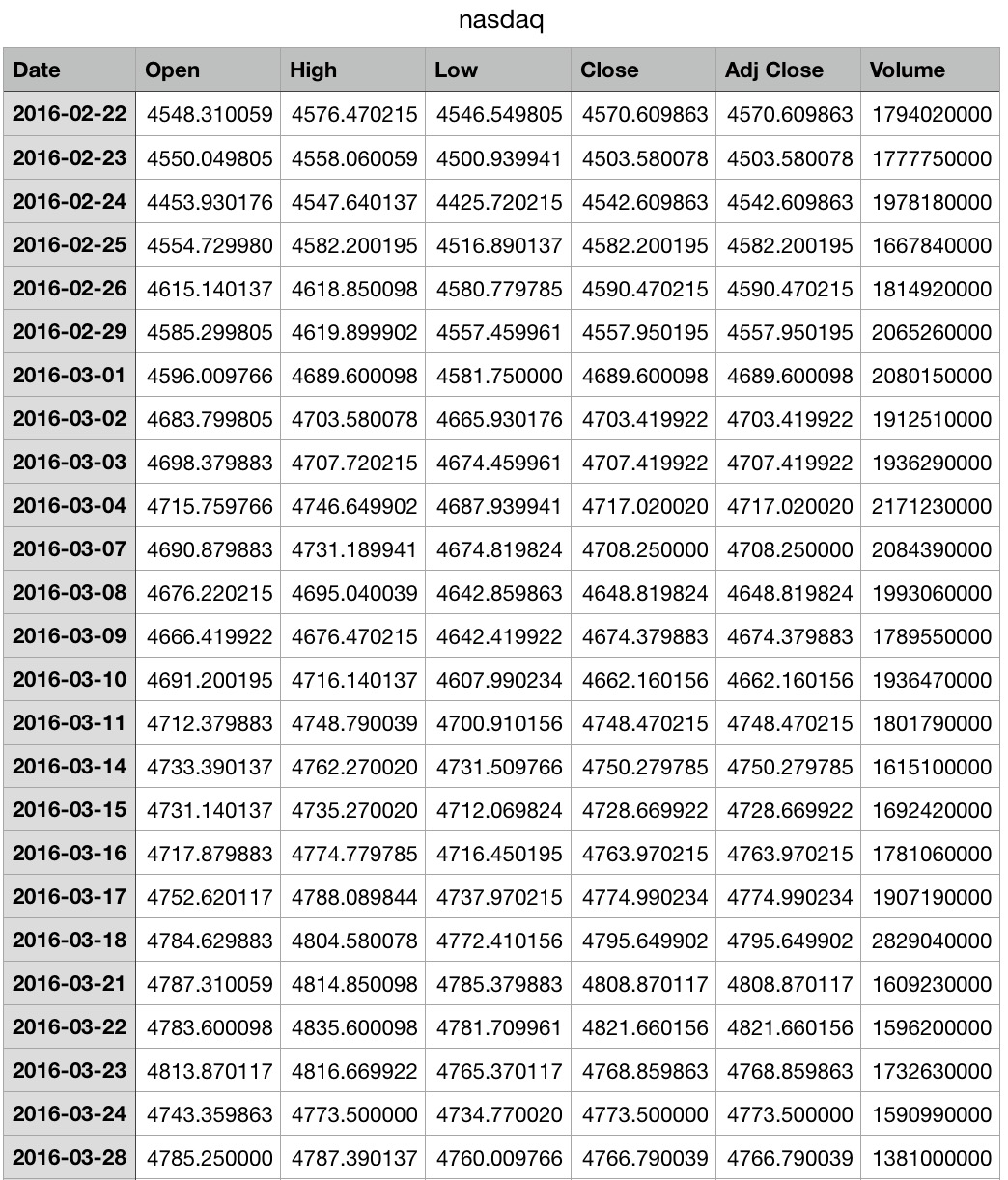
In order to answer these three questions, this project will use two statistical modeling method to do the research.

**Data sets**

There are three data sets will be used in this project, *nasdaq.csv*, *Nasdaq\_weekly\_return\_volatility\_detailed.csv* and *Nasdaq\_weekly\_return\_volatility.csv.*

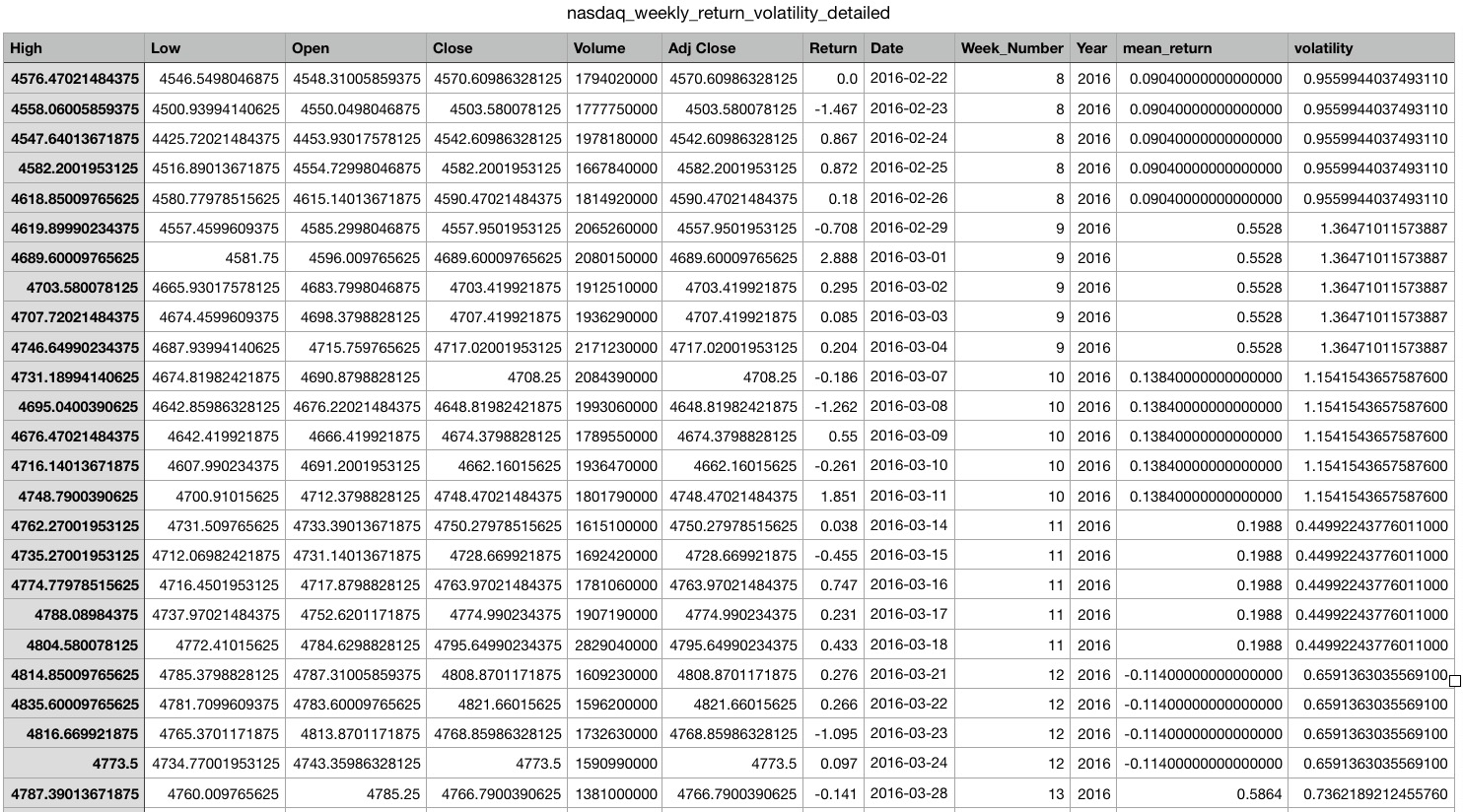
*Nasdaq.csv* contains 1260 rows and 7 columns. It stored the data of Nasdaq index from 2016-2-22 to 2021-2-22. The attributes are Date, Open, High, Low, Close, Adj Close, Volume.

***Nasdaq.csv***



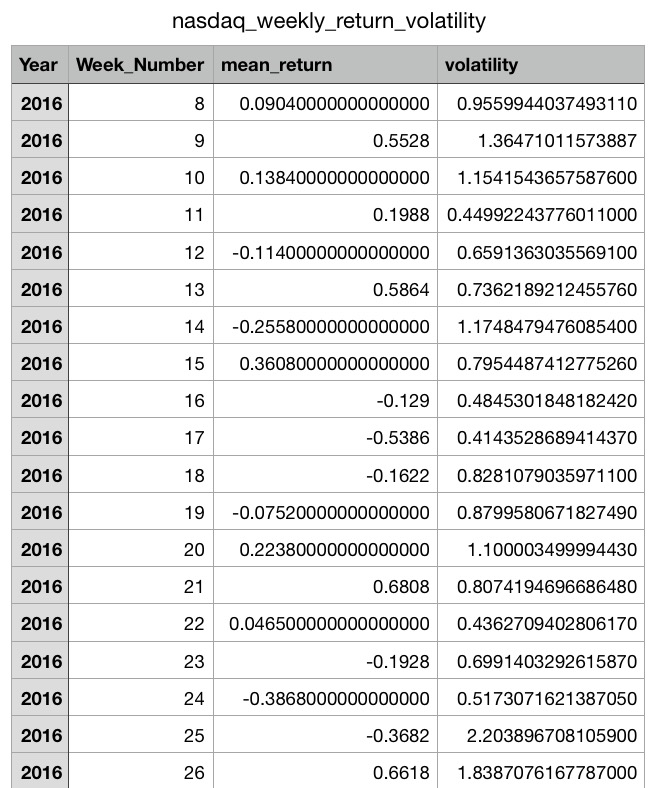
*nasdaq\_weekly\_return\_volatility\_detailed.csv* contains 1260 rows and 12 columns. It not only stored the data in *Nasdaq.csv*, but also have extra five attributes: Return, Week\_number, Year, Mean\_return, Volatility. This data set was generated by python code in order to show more details of the original data set.

***nasdaq\_weekly\_return\_volatility\_detailed.csv***



*nasdaq\_weekly\_return\_volatility.csv* is a short summary data set which contains 1260 rows but only 4 columns which are Year, Week\_number, Mean\_return, and volatility.

***nasdaq\_weekly\_return\_volatility.csv***



**Statistic Methods**

Two statistic methods will be used in this project: Simple and multiple linear regression. There are three questions, simple linear regression and multiple linear regression will be used for question one and two to determine the relationships between *Adj Close* and *Volume,* and *Volatility.*

**Project map**

* Preparing the data

1. Downloading *Nasdaq.csv* from Yahoo Finance
2. Run *compute\_stocks\_weekly\_return\_volatility.py*

* Data analyzing

1. Simple Linear Regression
2. Setting data

Using ***nasdaq.csv***

1. Scatter plot of the relationship
2. Regression line and equation
3. Correlation of the relationship
4. F-test
5. Multiple Linear Regression
6. Setting data

Using ***nasdaq\_weekly\_return\_volatility\_detailed.csv***

1. Correlation coefficients of variables
2. Pair diagrams
3. Multiple linear regression model
4. F-test

* Conclusion
* References

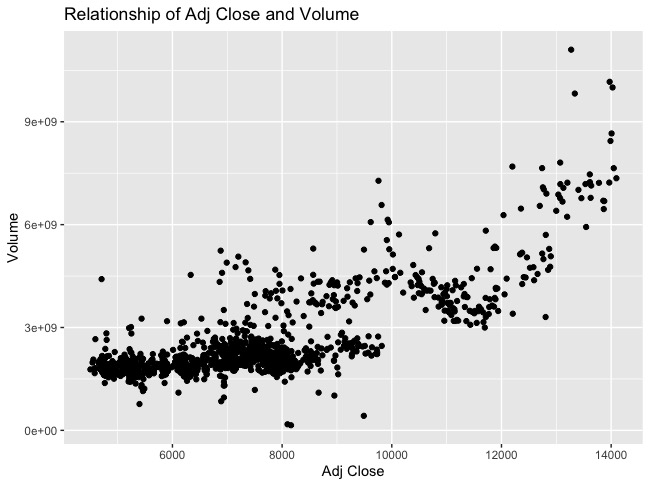
**Data analyzing**

1. Simple Linear Regression
2. Setting data

Importing CSV files of *Nasdaq\_weekly\_return\_volatility\_detailed.csv*, *Nasdaq\_weekly\_return\_volatility.csv, Nasdaq.csv* into RStudio.

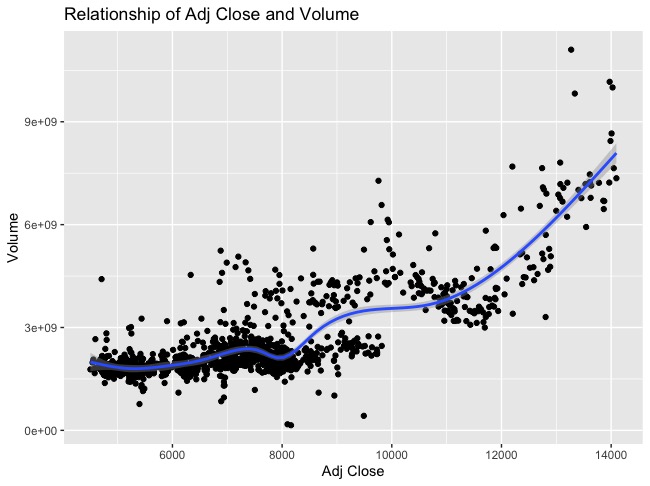
1. Scatter plot of the relationship between Adj Close and Volume

By using package “ggplot2”, building scatter plot of the relationship of Adj Close and Volume by using *Nasdaq.csv*.



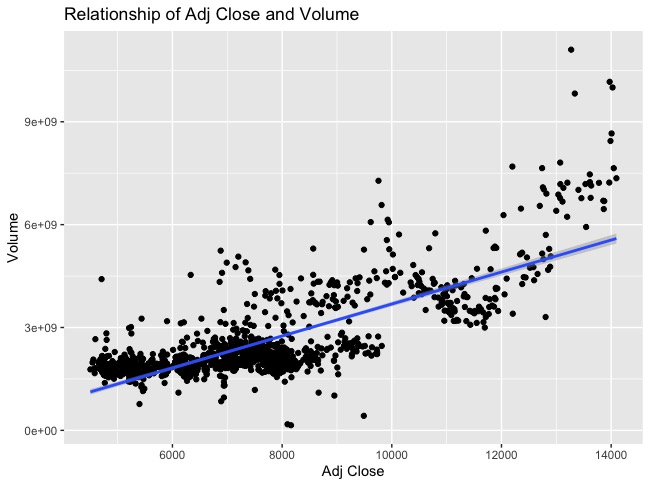
1. Regression line and equation

Using function geom\_smooth() to see the curve of the plot.



From above graph, the briefly summary is that with the increasing value of Adj Close, the volume of Nasdaq is increasing as well. Next, building regression line and find regression equation of the plot.

By adding method “lm” and formula y~x to the the function geom\_smooth(), the regression line displayed on the plot.



Using summary function to see the summary information of the relationship.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate Std. | Std. Error | T value | Pr |
| Intercept | -978889391 | 86214507 | -11.35 | <2e-16 |
| Adj Close | 466368 | 10946 | 42.61 | <2e-16 |

Residual standard error: 1334 on 1257 degrees of freedom

Multiple R-squared: 0.5909, Adjusted R-squared: 0.5905

F-statistic: 1815 on 1 and 1257 DF, p-value: < 2.2e-16

From above table, the least squared regression equation could be found, which is

**y = 466368x – 978889391**

slope parameter which is 466368 which is greater than 0 in this data set, it means that the explanatory variable increases. Greater value of Adj Close, bigger volume of Nasdaq stock will have.

1. Correlation of the relationship

Using cor() function to find the correlation coefficient of the relationship. Which gave the correlation coefficient is: **0.7686683.**

since the correlation coefficient is positive, so it means the scatter plot has positive association. Two variables are positive correlated. It also interprets that the strength of the relationship is strong.

1. F-test

In order to formally determine there is a linear relationship between Adj Close and Volume. F-test at α=0.05 will be used to formally test these two variables.

**Step 1:**

H0: β1=0 (there is no linear association)

H1: β1≠0 (there is a linear association)

α = 0.05

**Step 2:**

F = MS Reg / MS Res with 1 and n−3 degrees of freedom

**Step 3:**

Using R code, F1, 1257, 0.05 = 3.848867

Decision Rule: Reject H0 if F ≥ 3.848867

Otherwise, do not reject H0.

**Step 4:**

Create ANOVA table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq. | Mean Sq. | F value | Pr |
| Volume | 1 | 3.229e+09 | 3229009339 | 1815.2 | < 2.2e-16 |
| Residual | 1257 | 2.236e+09 | 1778843 |  |  |

F = 1815.2

**Step 5:**

Reject H0 since 1815.2 > 3.848867, We have significant evidence at the α=0.05 level that β1≠0. That is, there is evidence of a significant linear association between Adj Close and Volume.

To sum up, based on the tests above, the answer of question 1: what is the relationship between Adj Close and Volume is that the relationship is linear relation with least squared regression equation **y = 466368x – 978889391,** andcorrelation coefficient **0.7686683.** This means the strength of the relationship is strong. By F-test, since F value 1815.2 > 3.848867, linear association was formally confirmed between two variables.

1. Multiple linear regression
2. Setting data

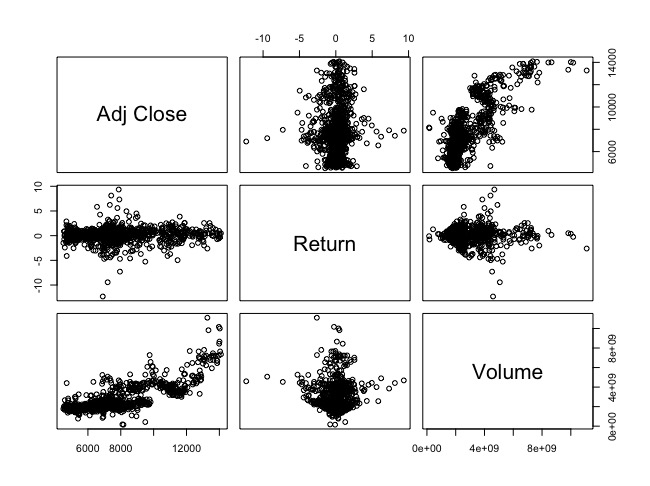
Using ***nasdaq\_weekly\_return\_volatility\_detailed.csv*** in this section to find multiple linear regression of Adj Close, Return, and Volume.

1. Correlation coefficients of variables

|  |  |  |  |
| --- | --- | --- | --- |
|  | Adj Close | Return | Volume |
| Adj Close | 1 | 0.05183447 | 0.76866835 |
| Return | 0.05183447 | 1 | -0.02845836 |
| Volume | 0.76866835 | -0.02845836 | 1 |

1. Pair diagrams

Pair diagrams of three variables by using R codes.



1. Multiple linear regression model

Before answering question 2, it is necessary to build a multiple linear regression model since there are three variables are considered this time.

By using R:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate Std. | Std. Error | T value | Pr |
| Intercept | -988930177 | 85794830 | -11.527 | < 2e-16 |
| Return | -65035606 | 17064100 | -3.811 | 0.000145 |
| Adj Close | 468522 | 10902 | 42.974 | < 2e-16 |

Residual standard error: 804900000 on 1256 degrees of freedom

Multiple R-squared: 0.5955, Adjusted R-squared: 0.5949

F-statistic: 924.6 on 2 and 1256 DF, p-value: < 2.2e-16

1. F-test

Perform an FF-test at the α=0.01 level to answer question2. Use the information from the ANOVA table below to help

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq. | Mean Sq | F value | Pr |
| Return | 1 | 1.6293e+18 | 1.6293e+18 | 2.5149 | 0.113 |
| Adj Close | 1 | 1.1964e+21 | 1.1964e+21 | 1846.7737 | <2e-16 |
| Residuals | 1256 | 8.1369e+20 | 6.4784e+17 |  |  |

**Step 1:**

H0: βreturn = βAdj Close= 0 (Return and Adj Close are not significant predictors of annual salary)

H1: βreturn ≠ 0 and/or βAdj Close ≠ 0 (at least one of the slope coefficients is different than 0; Return and/or Adj Close are significant predictors/is a significant predictor of Volume)

α=0.01

**Step 2:**

F= MS Reg / MS Res, df = 2, n-k-1

**Step 3:**

Using the software, Fk,n−k−1,α = F2,1256,0.01 = 6.655107.

Decision Rule:  Reject H0 if F ≥ 6.655107

Otherwise, do not reject H0

**Step 4:**

F = MS Reg / MS Res

MS Reg = 1.6293e18 + 1.1964e21 = 1.198e21

MS Res = 6.4784e17

So F = 1.198e21 / 6.4784e17 = 1849.267

**Step 5:**

Reject H0 since 1849.267 ≥ 6.655107.  We have significant evidence at the α=0.01 level that Return and Adj Close when taken together are significant predictors of Volume. That is, there is evidence of a linear association between Volume and Return and Adj Close.

In conclusion, it solved the question 2. The answer is that Adj Close and Return when considered together significant predictors of Volume.

**Conclusion**

From what the research displayed above, question 1: what is the relationship between *Adj Close* and *Volume* has been solved by simple linear regression. The answer is that the relationship is linear relation with least squared regression equation **y = 466368x – 978889391,** andcorrelation coefficient **0.7686683.** The strength of the relationship is strong based on the correlation coefficient. By F-test, since F value 1815.2 > 3.848867, linear association was formally confirmed between two variables. So, it proved the assumption, which was made by only looking the lines on the scatter plot.

Question 2: Are *Adj Close* and *Return* when considered together significant predictors of *Volume* has been solved as well by using multiple linear regression. The answer is that *Adj Close* and *Return* when considered together significant predictors of *Volume*. F statistic test showed 1849.267 ≥ 6.655107, that is a significant evidence at the α=0.01 level that *Return* and *Adj Close* when taken together are significant predictors of *Volume*. This is the evidence of a linear association between Volume and Return and Adj Close.

There are some limitations of this project. Firstly, linear regression assumed the regression line is a straight line not a curve, this is why the first figure showed a curve on the scatter plot. Since Nasdaq index is a real-world index, there are so many elements could have effect on Adj Close and Volume. So sometimes a straight regression line might not be correct. Secondly, especially in 2020, Nasdaq index experienced an historical phenomenon. The index dropped a lot in March 2020; however, because of the Fed's unlimited quantitative easing policy, an enormous number of cash was printed. As a result, the index was Miraculously recovered from May 2020 and created another historical high. So linear regression cannot present this especially term very clear. The reason why I choose historic data of Nasdaq because I tried to see how March 2020 looks like in statistic, nevertheless, the linear regression limited this, I will try another way to do the data analytics of Nasdaq index.

**Reference**

[https://finance.yahoo.com/quote/%5EIXIC?p=^IXIC&.tsrc=fin-srch](https://finance.yahoo.com/quote/%5EIXIC?p=%5eIXIC&.tsrc=fin-srch)

<https://onlinecampus.bu.edu/bbcswebdav/pid-8499014-dt-content-rid-48736472_1/courses/21sprgmetcs555_o1/course/module3/allpages.htm>

<https://onlinecampus.bu.edu/bbcswebdav/pid-8499015-dt-content-rid-48736520_1/courses/21sprgmetcs555_o1/course/module4/allpages.htm>

<https://onlinecampus.bu.edu/bbcswebdav/pid-8499017-dt-content-rid-48736599_1/courses/21sprgmetcs555_o1/course/module6/allpages.htm>