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?

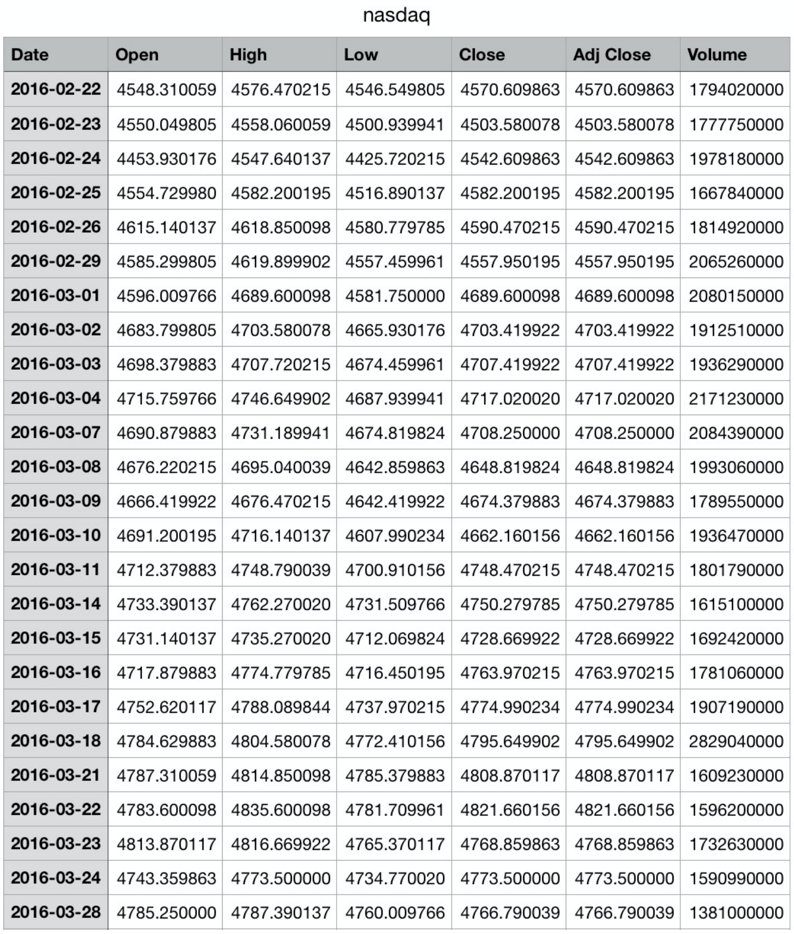
In order to answer these three questions, this project will use two statistical modeling method to do the research.

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

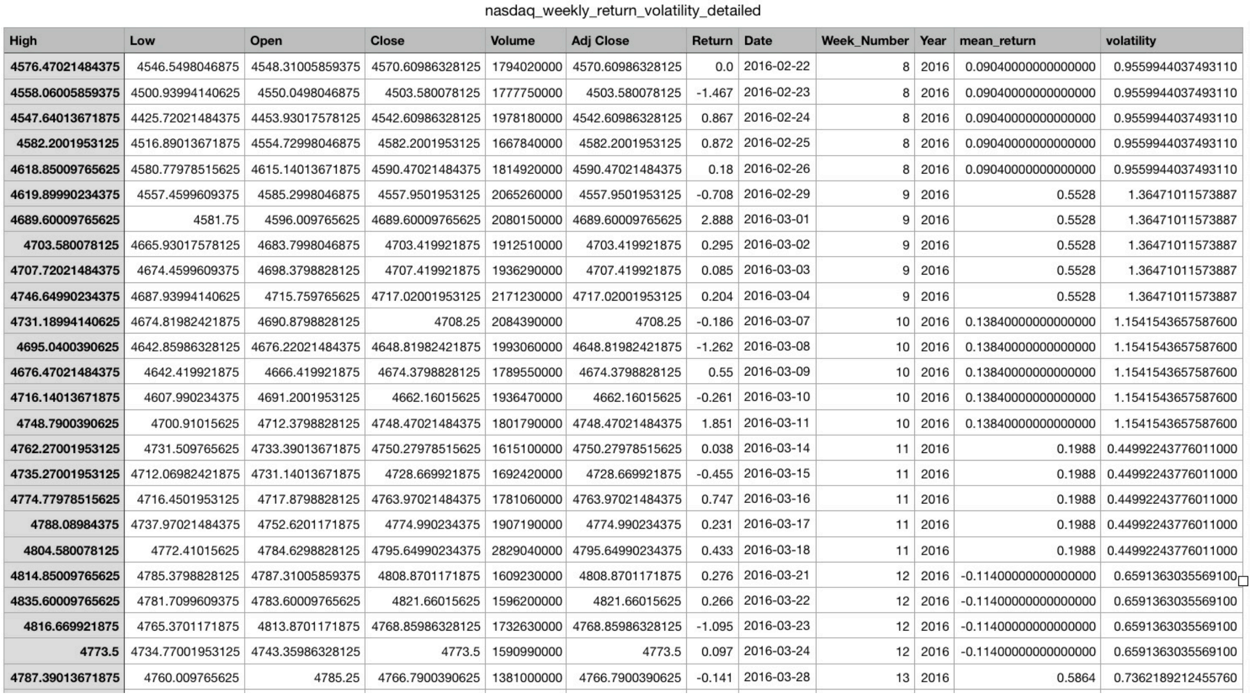


nasdaq

Date	Open	High	Low	Close	Adj Close	Volume
2016-02-22	4548.310059	4576.470215	4546.549805	4570.609863	4570.609863	1794020000
2016-02-23	4550.049805	4558.060059	4500.939941	4503.580078	4503.580078	1777750000
2016-02-24	4453.930176	4547.640137	4425.720215	4542.609863	4542.609863	1978180000
2016-02-25	4554.729980	4582.200195	4516.890137	4582.200195	4582.200195	1667840000
2016-02-26	4615.140137	4618.850098	4580.779785	4590.470215	4590.470215	1814920000
2016-02-29	4585.299805	4619.899902	4557.459961	4557.950195	4557.950195	2065260000
2016-03-01	4596.009766	4689.600098	4581.750000	4689.600098	4689.600098	2080150000
2016-03-02	4683.799805	4703.580078	4665.930176	4703.419922	4703.419922	1912510000
2016-03-03	4698.379883	4707.720215	4674.459961	4707.419922	4707.419922	1936290000
2016-03-04	4715.759766	4746.649902	4687.939941	4717.020020	4717.020020	2171230000
2016-03-07	4690.879883	4731.189941	4674.819824	4708.250000	4708.250000	2084390000
2016-03-08	4676.220215	4695.040039	4642.859863	4648.819824	4648.819824	1993060000
2016-03-09	4666.419922	4676.470215	4642.419922	4674.379883	4674.379883	1789550000
2016-03-10	4691.200195	4716.140137	4607.990234	4662.160156	4662.160156	1936470000
2016-03-11	4712.379883	4748.790039	4700.910156	4748.470215	4748.470215	1801790000
2016-03-14	4733.390137	4762.270020	4731.509766	4750.279785	4750.279785	1615100000
2016-03-15	4731.140137	4735.270020	4712.069824	4728.669922	4728.669922	1692420000
2016-03-16	4717.879883	4774.779785	4716.450195	4763.970215	4763.970215	1781060000
2016-03-17	4752.620117	4788.089844	4737.970215	4774.990234	4774.990234	1907190000
2016-03-18	4784.629883	4804.580078	4772.410156	4795.649902	4795.649902	2829040000
2016-03-21	4787.310059	4814.850098	4785.379883	4808.870117	4808.870117	1609230000
2016-03-22	4783.600098	4835.600098	4781.709961	4821.660156	4821.660156	1596200000
2016-03-23	4813.870117	4816.669922	4765.370117	4768.859863	4768.859863	1732630000
2016-03-24	4743.359863	4773.500000	4734.770020	4773.500000	4773.500000	1590990000
2016-03-28	4785.250000	4787.390137	4760.009766	4766.790039	4766.790039	1381000000

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

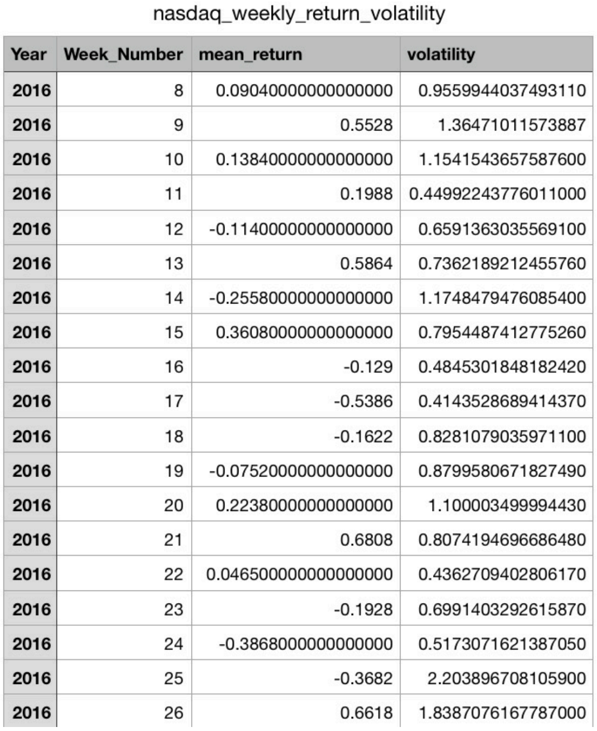


nasdaq_weekly_return_volatility_detailed

High	Low	Open	Close	Volume	Adj Close	Return	Date	Week_Number	Year	mean_return	volatility
4576.47021484375	4546.5498046875	4548.31005859375	4570.60986328125	1794020000	4570.60986328125	0.0	2016-02-22	8	2016	0.09040000000000000	0.9559944037493110
4558.06005859375	4500.93994140625	4550.0498046875	4503.580078125	1777750000	4503.580078125	-1.467	2016-02-23	8	2016	0.09040000000000000	0.9559944037493110
4547.64013671875	4425.72021484375	4453.93017578125	4542.60986328125	1978180000	4542.60986328125	0.867	2016-02-24	8	2016	0.09040000000000000	0.9559944037493110
4582.2001953125	4516.89013671875	4554.72998046875	4582.2001953125	1667840000	4582.2001953125	0.872	2016-02-25	8	2016	0.09040000000000000	0.9559944037493110
4618.85009765625	4580.77978515625	4615.14013671875	4590.47021484375	1814920000	4590.47021484375	0.18	2016-02-26	8	2016	0.09040000000000000	0.9559944037493110
4619.89990234375	4557.4599609375	4585.2998046875	4557.9501953125	2065260000	4557.9501953125	-0.708	2016-02-29	9	2016	0.5528	1.36471011573887
4689.60009765625	4581.75	4596.009765625	4689.60009765625	2080150000	4689.60009765625	2.888	2016-03-01	9	2016	0.5528	1.36471011573887
4703.580078125	4665.93017578125	4683.7998046875	4703.419921875	1912510000	4703.419921875	0.295	2016-03-02	9	2016	0.5528	1.36471011573887
4707.72021484375	4674.4599609375	4698.3798828125	4707.419921875	1936290000	4707.419921875	0.085	2016-03-03	9	2016	0.5528	1.36471011573887
4746.64990234375	4687.93994140625	4715.759765625	4717.02001953125	2171230000	4717.02001953125	0.204	2016-03-04	9	2016	0.5528	1.36471011573887
4731.18994140625	4674.81982421875	4690.8798828125	4708.25	2084390000	4708.25	-0.186	2016-03-07	10	2016	0.13840000000000000	1.1541543657587600
4695.0400390625	4642.85986328125	4676.22021484375	4648.81982421875	1993060000	4648.81982421875	-1.262	2016-03-08	10	2016	0.13840000000000000	1.1541543657587600
4676.47021484375	4642.419921875	4666.419921875	4674.3798828125	1789550000	4674.3798828125	0.55	2016-03-09	10	2016	0.13840000000000000	1.1541543657587600
4716.14013671875	4607.990234375	4691.2001953125	4662.16015625	1936470000	4662.16015625	-0.261	2016-03-10	10	2016	0.13840000000000000	1.1541543657587600
4748.7900390625	4700.91015625	4712.3798828125	4748.47021484375	1801790000	4748.47021484375	1.851	2016-03-11	10	2016	0.13840000000000000	1.1541543657587600
4762.27001953125	4731.509765625	4733.39013671875	4750.27978515625	1615100000	4750.27978515625	0.038	2016-03-14	11	2016	0.1988	0.44992243776011000
4735.27001953125	4712.06982421875	4731.14013671875	4728.669921875	1692420000	4728.669921875	-0.455	2016-03-15	11	2016	0.1988	0.44992243776011000
4774.77978515625	4716.4501953125	4717.8798828125	4763.97021484375	1781060000	4763.97021484375	0.747	2016-03-16	11	2016	0.1988	0.44992243776011000
4788.08984375	4737.97021484375	4752.6201171875	4774.990234375	1907190000	4774.990234375	0.231	2016-03-17	11	2016	0.1988	0.44992243776011000
4804.580078125	4772.41015625	4784.6298828125	4795.64990234375	2829040000	4795.64990234375	0.433	2016-03-18	11	2016	0.1988	0.44992243776011000
4814.85009765625	4785.3798828125	4787.31005859375	4808.8701171875	1609230000	4808.8701171875	0.276	2016-03-21	12	2016	-0.11400000000000000	0.6591363035569100
4835.60009765625	4781.7099609375	4783.60009765625	4821.66015625	1596200000	4821.66015625	0.266	2016-03-22	12	2016	-0.11400000000000000	0.6591363035569100
4816.669921875	4765.3701171875	4813.8701171875	4768.85986328125	1732630000	4768.85986328125	-1.095	2016-03-23	12	2016	-0.11400000000000000	0.6591363035569100
4773.5	4734.77001953125	4743.35986328125	4773.5	1590990000	4773.5	0.097	2016-03-24	12	2016	-0.11400000000000000	0.6591363035569100
4787.39013671875	4760.009765625	4785.25	4766.7900390625	1381000000	4766.7900390625	-0.141	2016-03-28	13	2016	0.5864	0.7362189212455760

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



nasdaq_weekly_return_volatility

Year	Week_Number	mean_return	volatility
2016	8	0.09040000000000000	0.9559944037493110
2016	9	0.5528	1.36471011573887
2016	10	0.13840000000000000	1.1541543657587600
2016	11	0.1988	0.44992243776011000
2016	12	-0.11400000000000000	0.6591363035569100
2016	13	0.5864	0.7362189212455760
2016	14	-0.25580000000000000	1.1748479476085400
2016	15	0.36080000000000000	0.7954487412775260
2016	16	-0.129	0.4845301848182420
2016	17	-0.5386	0.4143528689414370
2016	18	-0.1622	0.8281079035971100
2016	19	-0.07520000000000000	0.8799580671827490
2016	20	0.22380000000000000	1.100003499994430
2016	21	0.6808	0.8074194696686480
2016	22	0.046500000000000000	0.4362709402806170
2016	23	-0.1928	0.6991403292615870
2016	24	-0.3868000000000000	0.5173071621387050
2016	25	-0.3682	2.203896708105900
2016	26	0.6618	1.8387076167787000

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

  I. Simple Linear Regression
  1. Setting data
     Using ***nasdaq.csv***
  2. Scatter plot of the relationship
  3. Regression line and equation
  4. Correlation of the relationship
  5. F-test

  II. Multiple Linear Regression
  1. Setting data
     Using ***nasdaq_weekly_return_volatility_detailed.csv***
  2. Correlation coefficients of variables
  3. Pair diagrams
  4. Multiple linear regression model
  5. F-test

- Conclusion
- References

Data analyzing

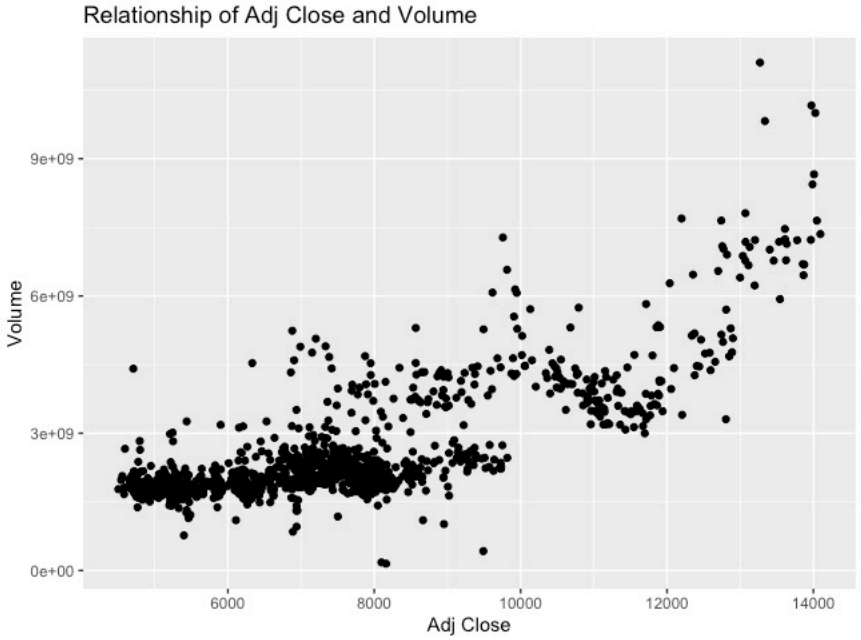
I. Simple Linear Regression

1. Setting data

Importing CSV files of *Nasdaq_weekly_return_volatility_detailed.csv*, *Nasdaq_weekly_return_volatility.csv, Nasdaq.csv* into RStudio.

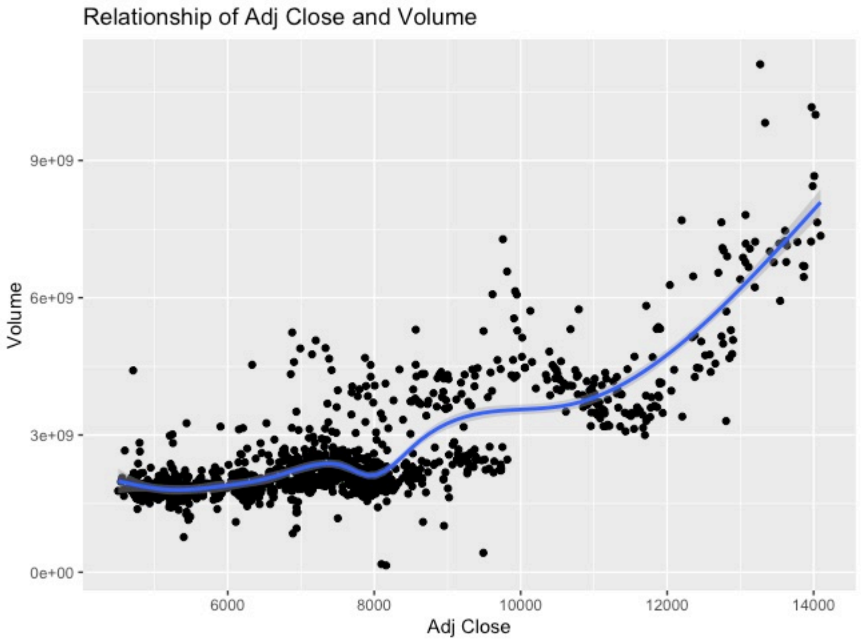
2. Scatter plot of the relationship between Adj Close and Volume

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



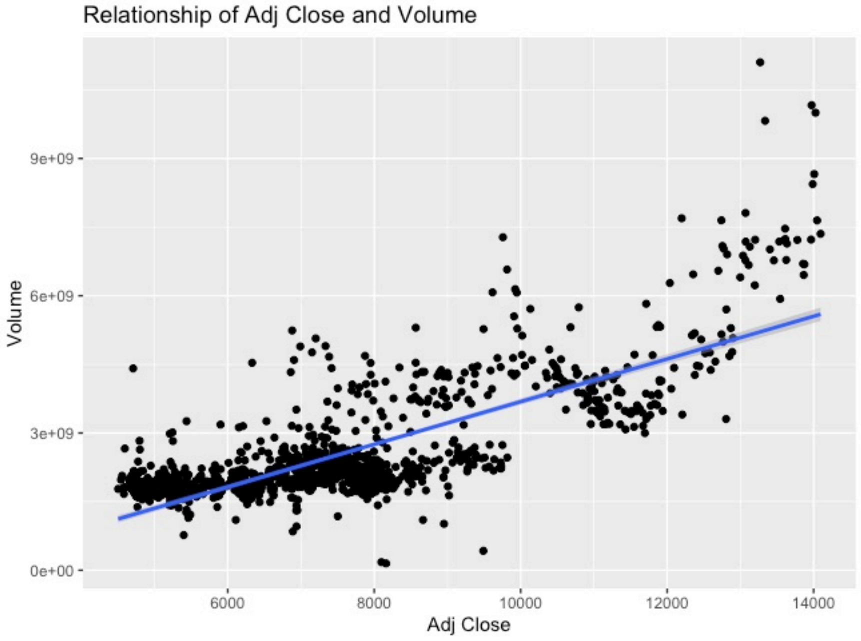
3. Regression line and equation

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.

4. 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.

5. 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:

H_0: $\beta_1=0$ (there is no linear association)

H_1: $\beta_1 \neq 0$ (there is a linear association)

$\alpha = 0.05$

Step 2:

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

Step 3:

Using R code, $F_{1,\,1257,\,0.05} = 3.848867$

Decision Rule: Reject H_0 if $F \geq 3.848867$

Otherwise, do not reject H_0.

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 H_0 since 1815.2 > 3.848867, We have significant evidence at the α=0.05 level that $\beta_1 \neq 0$. 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,** and correlation 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.

II. Multiple linear regression

1. Setting data

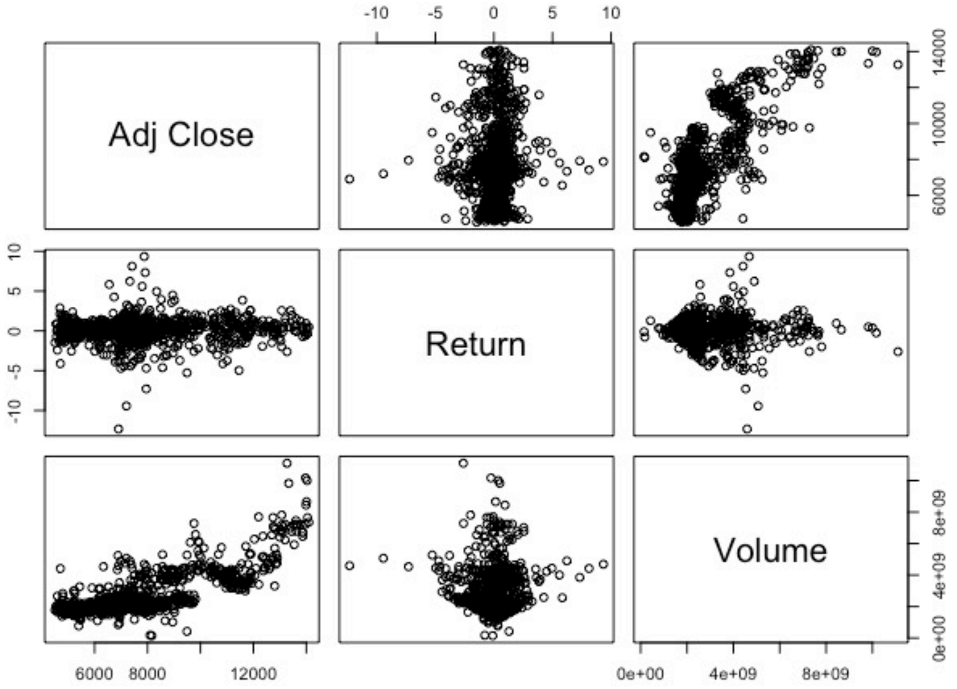
Using ***nasdaq_weekly_return_volatility_detailed.csv*** in this section to find multiple linear regression of Adj Close, Return, and Volume.

2. 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

3. Pair diagrams

Pair diagrams of three variables by using R codes.



4. 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

5. 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:

H_0: $\beta_{\text{return}} = \beta_{\text{Adj Close}} = 0$ (Return and Adj Close are not significant predictors of annual salary)

H_1: $\beta_{\text{return}} \neq 0$ and/or $\beta_{\text{Adj Close}} \neq 0$ (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, $F_{k,n-k-1,\alpha} = F_{2,1256,0.01} = 6.655107$.

Decision Rule: Reject H_0 if $F \geq 6.655107$

Otherwise, do not reject H_0

Step 4:

F = MS Reg / MS Res

MS Reg = $1.6293e^{18} + 1.1964e^{21} = 1.198e^{21}$

MS Res = $6.4784e^{17}$

So F = $1.198e^{21} / 6.4784e^{17} = 1849.267$

Step 5:

Reject H_0 since $1849.267 \geq 6.655107$. 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,** and correlation 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://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