UN3105 - Fall 2020
Final Project
Blinda Tian

I. Identify Question

A. Motivation

The coronavirus is one of the most influential global pandemics that has impacted the world in many aspects. My motivation comes from reading WSJ news's daily stock market summarizations that have always involved Covid updates since May. I believe that the stocks are driven by information diffusion, the market became more volatile since the spreading of covid news.

B. Objective

The main question of this study is how do vaccine news impact global investors' decision? Has the market been consistently reflected from COVID news? Typically, my focus of this study is vaccine news. As we gradually move towards the end of 2020, the need for vaccines is intense. With positive vaccine news, I assume that the global Investor reacts with positivity in the market.

Meanwhile, I noticed that in the country that Covid has been under control, for example, China, the reflection on Covid news in the stock market might be less. It would be interesting to compare the U.S and Chinese stock indexes to see if these two nation's investors respond the same way to vaccine news? Since investors' behavior can be detected from changes in the major stock market index, the main procedure would be examined inference between stock index and vaccine news.

II. Dataset

A. Ideal Dataset

Two sources are needed, one is stock indices that reflected on investors' sentiment, the other is market news that delivers vaccine news. Vaccine news is the proxy to detect investors' sentiments. Investors' behavior is gauged by reviewing the trading activity and direction of prices within a particular market. As for the role of media in the market, I believe that news events generally can cause price changes. This study will try to understand the role of vaccine articles on the stock index.

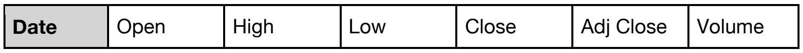
Chinese and U.S are comparable because these are two groups with significant distinctions in controlling Covid. China has managed to control the pandemic rapidly and effectively, whereas the US continues to be the country with the most cases.

If time and money are not a problem, an ideal stock index dataset is the three most widely followed indexes in the U.S., the S&P 500, Dow Jones Industrial Average, and Nasdaq Composite. Next, price movements are tracked should be in seconds with prior that the news all transmitted very quick in modern worlds. In the age of the internet, it takes seconds.

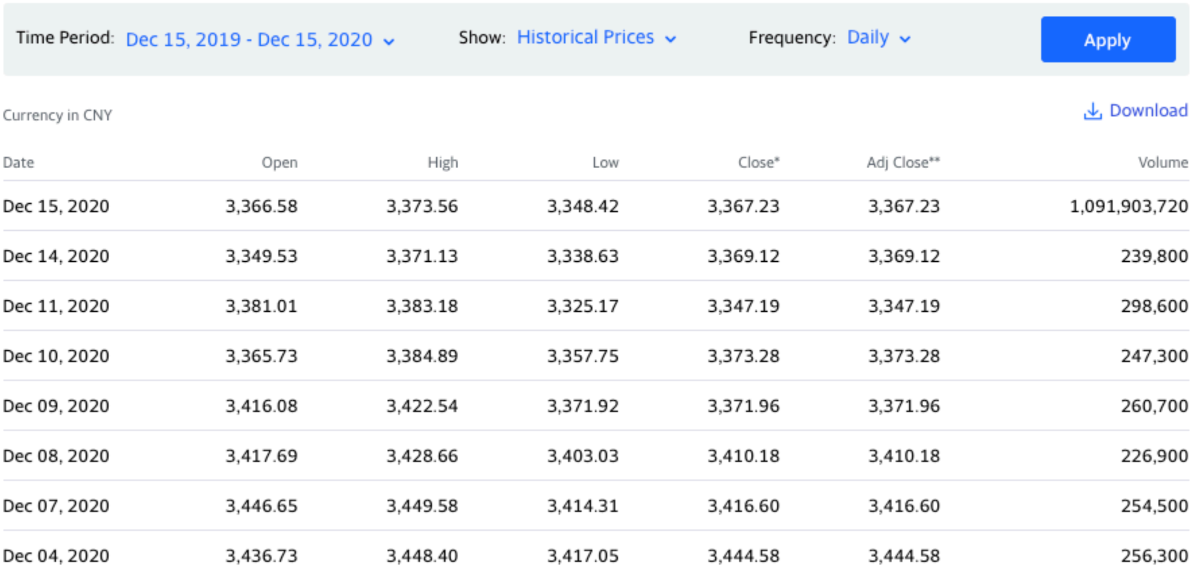
Regarding vaccine news, an ideal vaccine news dataset includes all-inclusive vaccine news from major financial publications that accessible to most investors. News sources are varied, included magazines, online news, trading platform, financial data vendor terminals. My assumption on news transmission is that major news should cover most of the mainstream vaccine news. However, there might be an attitude tendency in wordings or an appetite in news selection. Thus, the more major press I can cover, the more comprehensive I can capture news release.

B. Obtainable Dataset

Stock market :
Stock market price movements are extracted from Yahoo Finance. Yahoo Finance is a recognizable and reliable source that contains daily, weekly, and monthly prices for major indexes and individual stocks. It gives free access to an incredible amount of price information. Regarding country index, I choose S&P 500 to represent U.S stock investors' sentiments and SHE to be presented for the Chinese stock market. S&P is widely regarded as the best gauge of U.S equities since it comprised more stocks across all sectors compared to another index such as Dow's 30 and Nasdaq. Shanghai SE Composite Index is the most representative as it consists of all eligible stocks and CDRs listed on the Shanghai Stock Exchange. I've included a screenshot below of the S&P 500 directly from Yahoo Finance



Date	Open	High	Low	Close	Adj Close	Volume



Time Period: Dec 15, 2019 - Dec 15, 2020

Show: Historical Prices

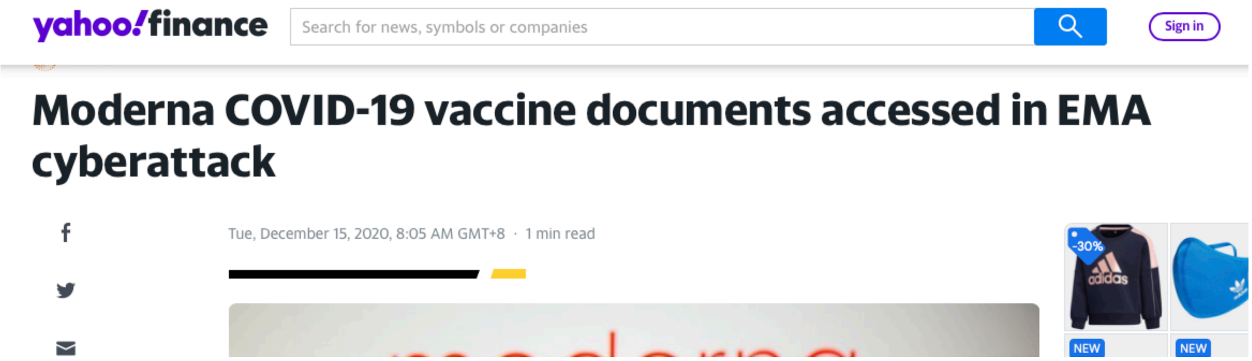
Frequency: Daily

Currency in CNY

Date	Open	High	Low	Close*	Adj Close**	Volume
Dec 15, 2020	3,366.58	3,373.56	3,348.42	3,367.23	3,367.23	1,091,903,720
Dec 14, 2020	3,349.53	3,371.13	3,338.63	3,369.12	3,369.12	239,800
Dec 11, 2020	3,381.01	3,383.18	3,325.17	3,347.19	3,347.19	298,600
Dec 10, 2020	3,365.73	3,384.89	3,357.75	3,373.28	3,373.28	247,300
Dec 09, 2020	3,416.08	3,422.54	3,371.92	3,371.96	3,371.96	260,700
Dec 08, 2020	3,417.69	3,428.66	3,403.03	3,410.18	3,410.18	226,900
Dec 07, 2020	3,446.65	3,449.58	3,414.31	3,416.60	3,416.60	254,500
Dec 04, 2020	3,436.73	3,448.40	3,417.05	3,444.58	3,444.58	256,300

Vaccine News:

Vaccines News are pulled from Yahoo Finance. Yahoo Finance is a widely used platform (free access) for investors to receive market news updates. Information includes headline and timestamp. Headlines, which have a similar length, are easier to parse and group than full articles, which vary in length. I can extract major sentiment easily and explicitly. The information below the updated vaccine news headlines and later, I will perform sentiment analysis on this information.



Moderna COVID-19 vaccine documents accessed in EMA cyberattack

Tue, December 15, 2020, 8:05 AM GMT+8 · 1 min read

Time:
Given time strain and the accessibility, I decided to study stock movement in a short time span – September. All stock price level and vaccine news are extracted from this month to get a consistent sample.

SEPTEMBER 2020

Sun	Mon	Tue	Wed	Thu	Fri	Sat
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	1	2	3

C. Differences

Firstly, this sample frame is conducted from data in September. It is also worth to notice that because of the limit in the time frame, the ability to extrapolate any inferences beyond our sample is one of the concerns. The sample dataset is still relevant to test out the main question given a specific time frame. Thought it does not cover holistic sentiments and behavior of stock market movement across a long period. The data can imply what happened in September.

Secondly, vaccine news is grabbed manually, which covers most of the vaccine news. The inclusiveness of vaccine news does not measure. Besides, each press has its tendency to depict events. For example, the fact that articles are written differently could provide different perspectives on the same story. Plus, the usage of words is varied, and investors observe and perceive them differently. However, a consistent textual pattern from one news platform will improve the sentiment analysis.

Thirdly, the relationship between financial news article release and its impact on stock price within minutes of release has not been fully studied. Instead, my study will focus on the closing price with the expectation that investors have taken in vaccines news and response to it in one trading day.

Fourthly, The inherent characteristic of an investor, such as geographical positions, investment portfolios, risk preference, etc. Thus, this project will study the majority of investors' responses regarding vaccine news. A general investment trend can shed some light on investors' sentiments.

Fifthly, Vaccines news's authenticity is not investigated. Furthermore, some news "events" are not vaccine specific. These include government action on the vaccine, analyst rating changes, politicians' response to vaccine distribution. The level of severity is not covered here.

III. Examined the Dataset (10)

A. Vaccine News Data

I manually grab daily news with the keyword "vaccine" from Yahoo Finance and forming a news dataset that consists of all news headlines and timestamps. News is released from 2020.09.01 to 2020.09.30. Then, I am sorting the news by chronological ascending order. 56 news related to vaccines has been archived.

Sentiment Features/Context Analysis

The following introduces the algorithm I used to programmatically predicting sentiment from news headlines. The ultimate goal is to categorize vaccine news into three sentiments that correspond to public opinions on Covid and labeled news as "positive" " negative" "neutral". Neutral essentially meaning no news. Since the Higher the emotional vocabs, the more informative. Thus, I use the most common sentiment words to analyze headlines. The underlying process is to match headline words with high and emotionally detected lexicons. Thus, a new Data frame was created with each headline, timestamp, and respective sentiment categories.

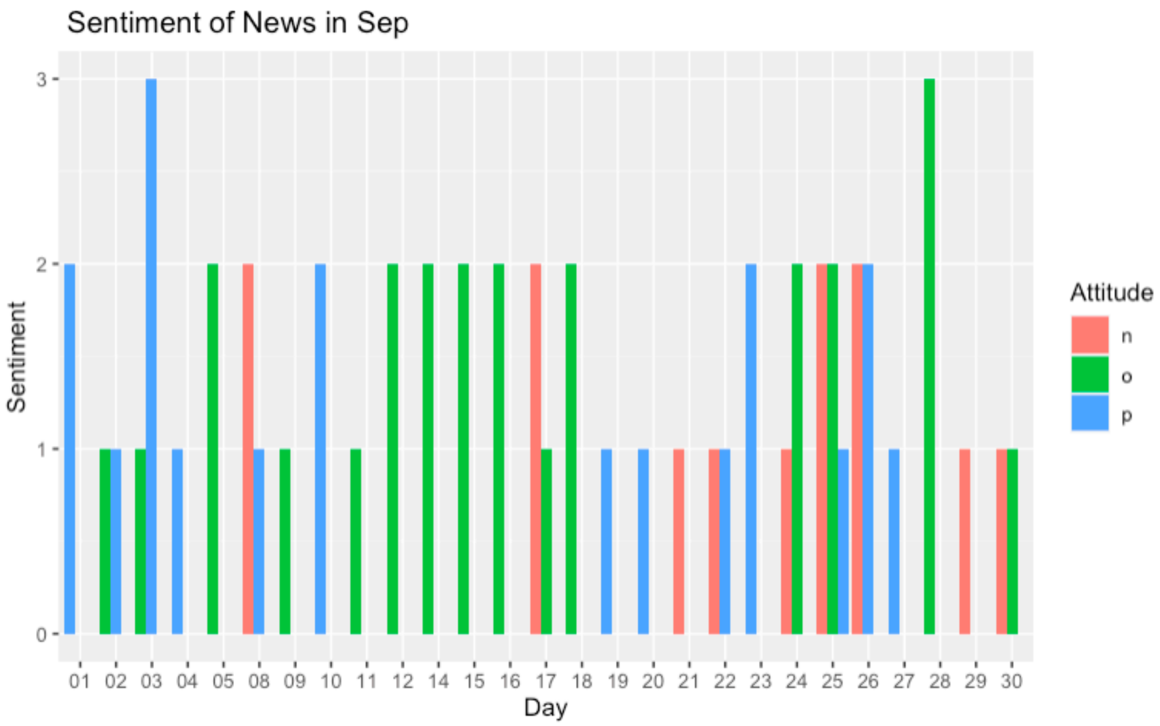
Step 1	Text cleaning: eliminate questions mark, remove numbers, break headlines into single words
Step 2	Check frequency of usage and unique characters
Step 3	extract bag-of-words features from the text; Used three different lexicons to discern sentiments, which all are provided relatively similar results. afinnsentiments bingsentiments loughransentiments Finally, I choose "bingsentiments" that includes larger vocabularies (6,776 words) and directly pass in positive and negative categories. This match has given a number of 25 vocabs, which is not enough. Thus, I also add some words' sentiments based on my finance knowledge intuition.

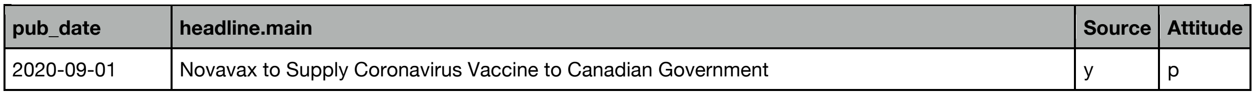
Step 4	Omitted those without delivering great emotions. My assumption is that investors would not respond intensively to the news. Thus, that would identity which day is “strong sentiment news day” and “no news day
Step 5	Based on the sentiment graph, assign one final major sentiment on one single trading day. That involves dropping Vaccine News data with missing values because the stock markets data is not available for weekends or national holidays.

Package used	library(textdata)
	library(tidytext)
	library(tidyverse)
	library(stringr)
	library(tm)
	library(magrittr)

Concern	The subjectivity on headline words; And the difference in each individual perceive the headline is not considerate here. Thus, the sentiment result may have bias on assigning sentiments and affect whether or not investors would response to news or not.

Finally, the news data is ready to be manipulated and analyzed with stock index. The following table is a screenshot on news sentiment data. Then, I also plot all sentiment results into a plot over time. This is a visualization on single day with different color represent the sentiment categories of vaccine news.





pub_date	headline.main	Source	Attitude
2020-09-01	Novavax to Supply Coronavirus Vaccine to Canadian Government	y	p

2020-09-01	AstraZeneca to launch late stage COVID-19 vaccine trial	y	p
2020-09-02	AstraZeneca Boosts Manufacturing of Its Coronavirus Vaccine With a $66 Million-Plus Deal	y	p
2020-09-02	Trump Is Winning the Vaccine Debate With Public-Health Experts	y	o
2020-09-03	CDC Tells States to Get Ready for Nov. 1 Vaccine Distribution	y	o
2020-09-03	Be Ready To Distribute A COVID-19 Vaccine In Two Months, CDC Tells All States	y	p
2020-09-03	Merck CEO sees human trials for COVID-19 vaccine candidate 'fairly soon'	y	p
2020-09-03	Novavax Analyst Says Coronavirus Vaccine 'Promising': 6 Takeaways From Phase 1 Results	y	p
2020-09-04	Coronavirus update: CDC, Fauci optimism fan vaccine hopes; GSK, Sanofi start Phase 1 trials	y	p

	word	**sentiment**
1	fairly	positive
2	optimism	positive
7	eases	positive
8	ready	positive
9	better	positive
10	boost	positive
11	best	positive
12	guidance	positive
14	approval	positive
15	promise	positive
18	trust	positive
19	protect	positive
20	work	positive
22	tops	positive
24	safe	positive
28	Supply	positive

B. Stock Data

I started my sample construction by extracting historical daily stock market data from Yahoo.com. Data starts from Sep 1st, 2020 and ends on Sep 30th, 2020. The daily open-high-low-close (OHLC) prices and volumes of two have been collected. S&P covers 22 trading days and SSE covers 21 trading days in September. I applied several filters to refine the data.

Anomalies	Notice that the length of SSE trading days is 22, and U.S S&P 500 index has 21 days. That is because of September 7th 2020 is Labor Day. Market close in U.S. Thus, I decided to omit September 7th 2020 historical price in SSE, so that I can merge these two datasets are comparable
Stock index autocorrelation:	**acf(sse_new$Adj.Close)** seen figure below Seen autocorrelation in stock indedx. The values (front lag) beyond blue line which the autocorrelations are statistically significantly different from zero. Meaning that the present value is depend on past stock price. Thus, I normalized data using diff(); return a suitable lagged differences . **acf(diff(sse_new$Adj.Close))** seen below; Then the acf graph looks less correlated. Also, since I only investigate in September, a temporarily very close and similar period. From the autocorrelation graph shown below, S&P and SSE perform similar pattern. Thus, the time series issue can be ignored if I applied this into a difference in difference model.
Correlation between two stock index	Seen graph below. This has validated a genuine positive and moderate relationship between S&P 500 and SSE market return.
Time Zone; One day Lag	There is a 12-hour time zone difference between the U.S and China. Since Yahoo finance's press release will follow by U.S time. Thus, I try to validate which period I should cover in China responding to Vaccine news by implementing a correlation test in one day lag on the Chinses market. The result being there is a higher correlation between the S&P500 index at day t and the SSE index at t+1. Thus, I decided to use the one-day lag SSE dataset going forward.

Package used	library("lubridate"); library("ggplot2"); library(tidyverse); library("lubridate");library(xts); library(timetk); library(PerformanceAnalytics); library(timetk)

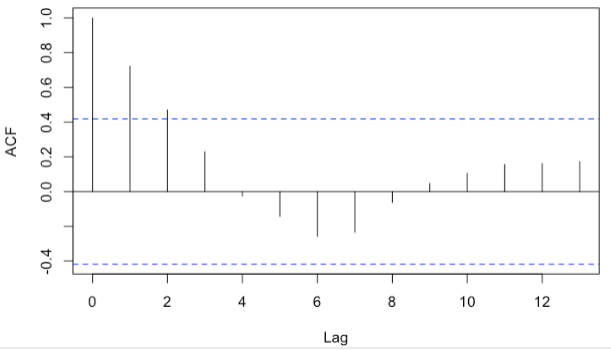


Figure: acf(sse$Adj.Close)

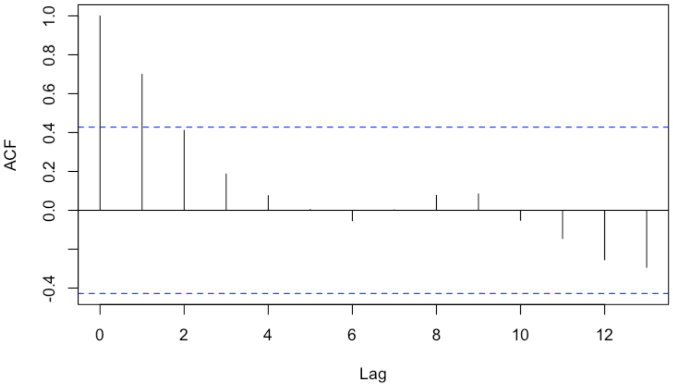


Figure: acf(snp$Adj.Close)

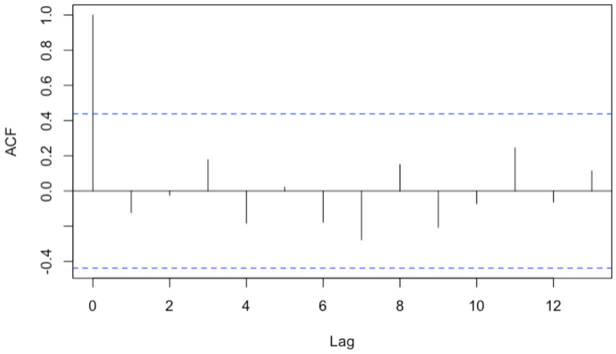
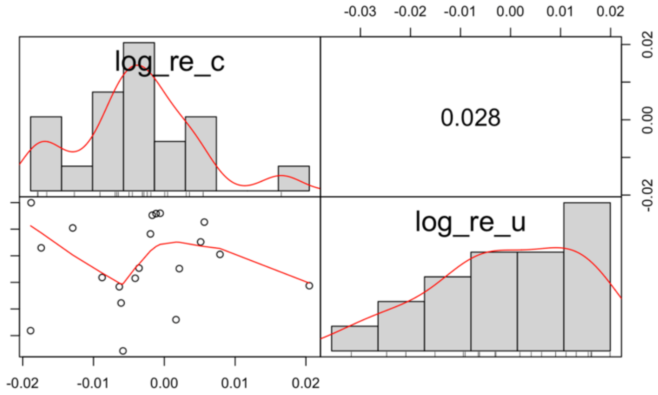
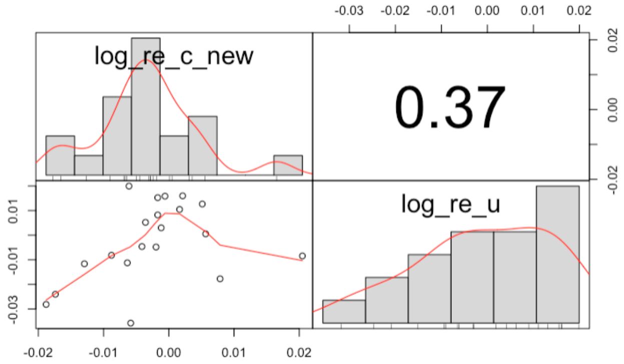


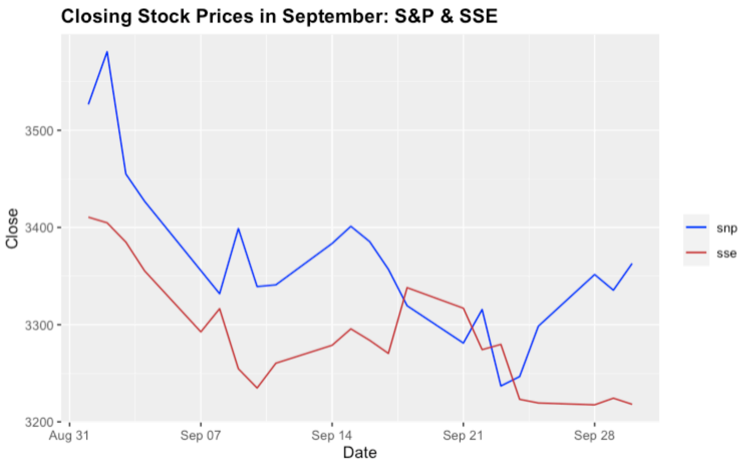
Figure: acf(diff(sse_new$Adj.Close))



Correlation between S&P and SSE



Correlation between S&P and SSE (one day lag)



C. Summarization

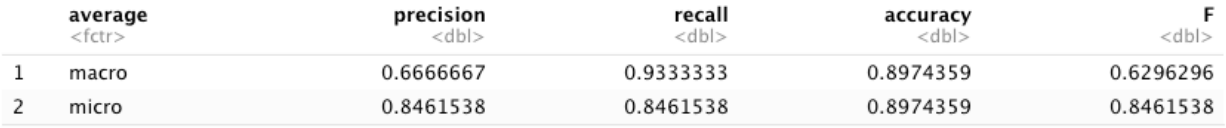
Number of Vaccine articles	57
Number of Positive vaccine news	19
Number of Negative vaccine news	13
SSE trading days	22
S&P trading days	21

Stock index Return							
	Min	1st Qu.	Median	Mean	3rd Qu	Max.	SD
SSE	-0.018886	-0.006965	-0.002766	-0.003266	0.001768	0.020476	0.009754376
S&P 500	-0.035758	-0.011344	-0.002048	-0.002376	0.011013	0.019945	0.01446959

Notice that there is a higher SD in S&P 500 index in Sep, suggesting that S&P 500 spread out far from the mean, implies more volatility involved in U.S market

D. Validation

To validate the preciseness on the sentiment algorithm. I randomly selected 20% form the news dataset. Then run a simple stimulation seeing how this algorithm does will be different from my own judgment. The result is promising. But still, this is a comparison to my own personal view on news.



	average <fctr>	precision <dbl>	recall <dbl>	accuracy <dbl>	F <dbl>
1	macro	0.6666667	0.9333333	0.8974359	0.6296296
2	micro	0.8461538	0.8461538	0.8974359	0.8461538

IV. Analysis (15 pts)

A. Overview

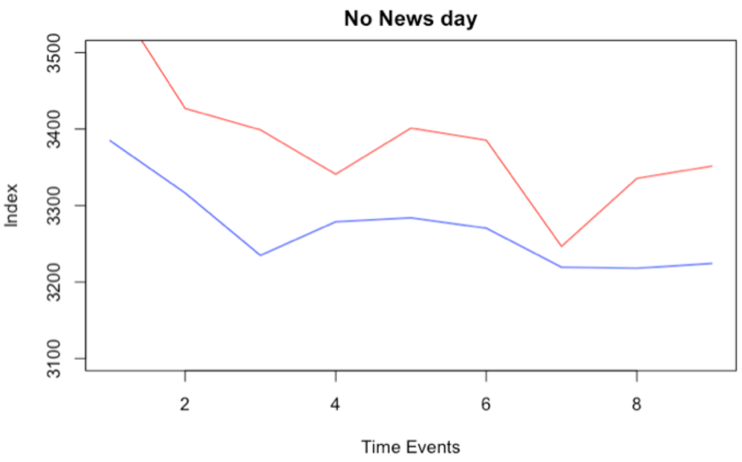
Goal
From my data, I would wish to know whether U.S investors that observe vaccine news will an increase in buying than Chinese investors; Therefore, I need to test whether the time effect is different when vaccine news intervene compared with when they don’t.

Model - Difference in Difference
Data is structured as Panel Data. I will use Difference in Difference to analyze the data. This model design can study the differential effect of a treatment on a 'treatment group' versus a 'control group’. The reason involved that DID model allowed to omit unobserved or unmeasured variables, thus controlling for omitted variable bias. Since there are many factors that impact on stock price, DID can single out the effect of vaccine news. Also, since I am interested in two comparison group level data. DID can compare groups that have different levels of characters and outcomes.

My assumption is that U.S will response intensely (maybe 30 more points increase in S&P500 index) on vaccine news.

B. Assumption

Parallel Trend Assumption
Parallel trend assumption requires the difference between the ‘treatment’ and ‘control’ group is constant over time in the absence of treatment. The following visual inspection is useful. Violation of parallel trend assumption will lead to biased estimation of the causal effect. Although. The following graph showing a general parallel except on part. The reason might be a confounder pop up in U.S market (red line) to drive the abnormalities; Thus, I decide to exclude this from the analysis.

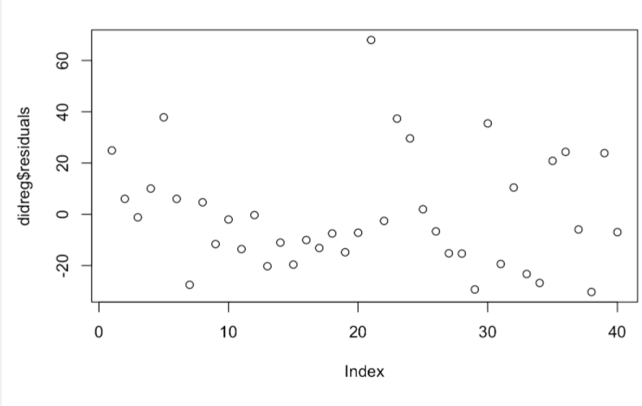


Linearity – valid	The regression model is linear in parameters	dummy variables have met the assumption of linearity
Mean of residual is zero – valid	mean of residuals is approximately zero, this assumption holds true for this model.	mean(didreg$residuals) [1] 7.544204e-16
Homoscedasticity of residuals or equal variance – may violated	residuals are spread widely; thought there is no pattern in magnitude, but seemed most of residuals fall below zero line.	Plot(model&residual)

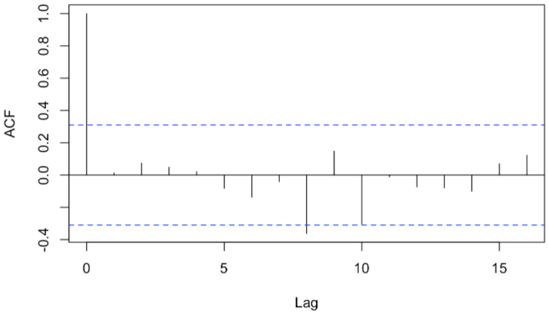
Autocorrelation – valid	Suggesting that there is not a clear correlation across time within entities; beside there is a abnormality in the front leg	acf(didreg$residuals)
		lmtest::dwtest(didreg)

No perfect multicollinearity	VIF for an Xvariable should be less than 4 in order to be accepted as not causing multi-collinearity	vif(didreg) News Country_treat News:Country_treat 1.856402 2.169670 2.984029
Normality of residuals – violated (solved)	Approximately follow the; three outlier exist; The presence of outliers may affect the interpretation of the model because it	Seen graph; solved by removing outliers; After that, the overall SE is smaller

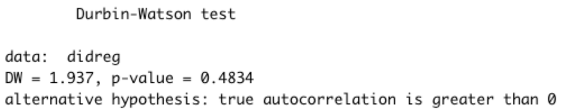
	increases the a relative standard error	
Packagd used	library(car)	



Plot(model&residual)

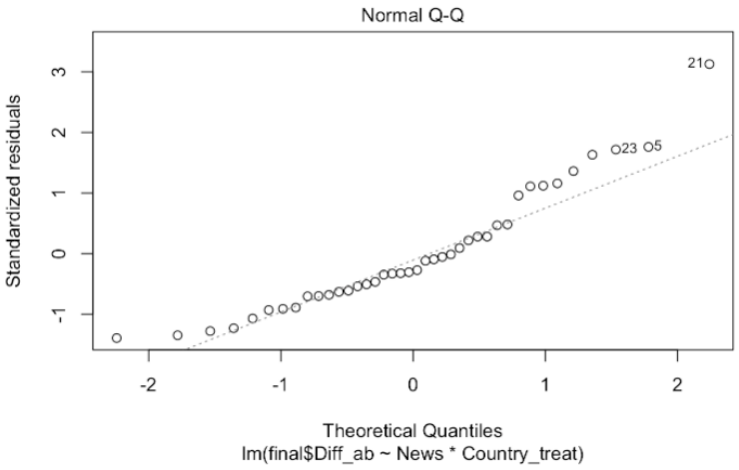


acf(didreg$residuals)



```
	Durbin-Watson test

data:  didreg
DW = 1.937, p-value = 0.4834
alternative hypothesis: true autocorrelation is greater than 0
```



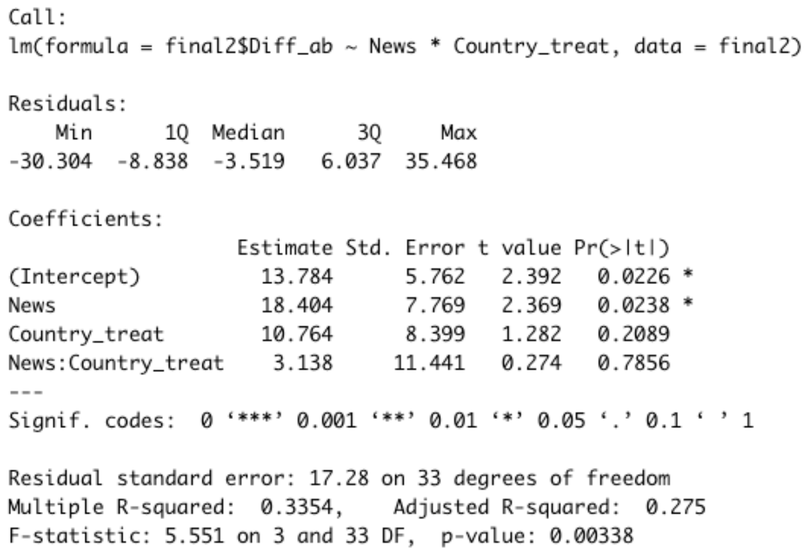
plot(model, 2)

C. Model

$$\text{Index} = \beta_0 + \delta_1\text{News} + \beta_1\text{Country} + \gamma(\text{News} \times \text{Country}) + e \quad (1)$$

Dependent	Stock index daily close level (diff normalized), which is a continuous variable
Dummy News Treatment	Time is binary (0,1) where 0 is no vaccine news and 1 was days that have vaccine news.
Country Treatment	Country is also binary, where 0 is a Chinses stock index and 1 is U.S stock index
Interaction	Create an interaction between treated and country
Formula	lm(final$Diff_ab ~ News*Country_treat, data = final)

D. Result



```
Call:
lm(formula = final2$Diff_ab ~ News * Country_treat, data = final2)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.304  -8.838  -3.519   6.037  35.468 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)  
(Intercept)          13.784      5.762   2.392   0.0226 *
News                 18.404      7.769   2.369   0.0238 *
Country_treat        10.764      8.399   1.282   0.2089  
News:Country_treat    3.138     11.441   0.274   0.7856  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.28 on 33 degrees of freedom
Multiple R-squared:  0.3354,	Adjusted R-squared:  0.275 
F-statistic: 5.551 on 3 and 33 DF,  p-value: 0.00338
```

Country_treat estimator is the estimated mean difference in stock index between US and China prior to the intervention. That there is 10 points difference in daily changed of stock index.

News dummy variables is statistically significant at 5%. This is the expected mean of change in stock index from no news to after exposure to news intervention in U.S. It implies that news could generate a 18 points changes in S&P 500.

β3 is the difference in differences estimator. It tells us the expected mean change in stock index from before to after was different in the two groups. Though the result is not statistically significant, the sign implies a positive difference between U.S and Chinese investors responding to vaccine news.

Overall, U.S investors reacts actively in stock market compare to Chinese investors. That explains the large standard deviations in S&P index summarization.

E. Concerns:

Model estimator insignificant: The model does not have interaction effect significant, perhaps there is an absence of evidence of another interaction effect.

V. Conclusion

The goal of this study is to infer how Chinese and U.S investors respond to Covid Vaccine news. The study has broken into two stages. One is a sentiment processing technique to understand the emotion behind the headlines. Then, the model I present above links the vaccine messages with the U.S and Chinese major indexes, respectively.

The result matches my expectation. Generally speaking, findings suggest that U.S stock investors react strongly to vaccine news and see a larger change responding to the headlines. On the other side, Chinese investors seemed to be more rational to interpret the news. One explanation might be the need for a vaccine is less urgent in the U.S compared to China.

A further thing I can explore is to see how investors respond does differently to vaccine news content, including a positive, negative, neutral one. Also, it will be great to compare reactions in different months to see do investors express an intense need for a vaccine back in June and now in December. I also want to learn how statisticians address a not perfect parallel trend in DID model.

Github Code: https://github.com/BlindaTian/FinalStats