Secondary Inquiry Report

Work Group 1

November 20, 2020

Question

Refined question

The broad question of this project is how does the COVID-19 pandemic impact people's emotion. There are two key terms in our broad question. The first one is the COVID-19 pandemic. Among various aspects of the pandemic, we primarily focus on the total number of confirmed cases and deaths. The second key term is emotion. Since it's difficult to define emotion scientifically, we are currently looking at people's emotion through their tweets in Twitter. Unfortunately, the dataset we found only have tweets with `#covid19` hashtag. Therefore, we only look at the tweets with `#covid19` hashtag. To sum up, our refined statistical question is how does the total number of infections and deaths affect the sentiment of tweets with `#covid19` hashtag in the US. Here we only focus on the US in order to narrow down our research topic.

Applicability

The major limitation of our refined question is the lack of accuracy in measuring emotion. Although Twitter is one of the most popular social media platforms in the US, the tweets may not be able to reflect people's emotion comprehensively. For example, people who have extreme negative attitudes might not want to post their real thoughts on Twitter. Also, there are lots of homeless people, elderly people and patients who do not use Twitter at all. Therefore, our tweets-based emotion analysis might be biased. However, we still believe that our method would produce compelling results because there is no perfect way to synthesize human emotion.

Considering only the tweets with `#covid19` hashtag adds limitation to our research as well. There certainly are many COVID related tweets that don't have `#covid19` hashtag. Fortunately, we recently found a website containing streaming tweets without using any specific query. We haven't got time to work with the new data yet. For the next step, we can extract random samples from the streaming tweets to replace the datasets we have right now. However, we are also aware that the new data might add more noise to our models because it contains tweets that are not related to COVID. Therefore, we need to do more research about this in the future.

Another major limitation of our refined question is the way we quantify the severity of COVID. The number of total infections and deaths is not the only thing that concerns people. Some people might have negative attitudes because they have to stay at home. Others might be angry about losing their jobs. Therefore, in the future, we can look for variables that explain other aspects of the pandemic.

Data Set

Description and Criticism

Our dataset includes tweets, sentiment and regional cases.

Tweets: It's collected with a hashtag of `#covid19` from July 25 to August 30 in 2020, including tweet texts, the location and the time that tweets were published. The dataset contains almost 190,000 lines. Due to the uneven sample size in different countries, in order to avoid too small sample size in some countries, we only select samples from the United States and then match them with the number of COVID cases according to the states in the 'Regional cases' dataset. The limitations are shown as below:

- The time span is short. If the COVID continues to be stable (steadily severe or increasing), changes of people's sentiment may not be remarkable and the underlying model can be no relationship.
- The dataset is collected with a hashtag `#covid19`. Users might not tag it when they make complaints, especially in the extreme emotion. Those who tend to tag may want to publish news or comments, which might cause the sentiment we analyses biased.
- The texts of the tweets are not completely recorded. Some of the tweets are followed with apostrophe and don't contain all the massage, which might cause some deviations from the original tweets, inducing the incorrectness of the outcome of our sentiment analysis.

Sentiment: A relatively small size of tweets are pulled from Twitter and manually tagged with "Negative", "Extremely Negative", "Positive", "Extremely Positive" and "Neutral". It is treated as a test dataset, to validate and assess the sentiment analysis method we use.

Regional cases: It collects information from 50 US states, the District of Columbia, and 5 other US territories and provides the testing data of positive and negative results, pending tests, as well as total hospitalizations, deaths and recovered. We combine this dataset with the "Tweets" to analyze whether the death rate, number of the confirmed cases, rate of the deaths and recovered, etc. are associated with the users' moods.

Pre-Process:

Location Cleaning: Since the messy location information in the data set holds no fixed format, and there are even some typo in the location information, then it is necessary to do the text cleaning. The main idea of location cleaning is to transform to lower cases, search for the location string containing the some special substring, like "usa", the state name or city name, and then extract or convert to the corresponding state name.

Text Cleaning:

Sentiment Analysis: We achieve the sentiment analysis whereby the 'bing' lexicon. The main idea is to evaluate the sentiment (score) of each word, based on the lexicon, and consider the sum as the sentiment of the whole text. Then we omit the tweets with equal amount of positive and negative words, and then label the tweets with more positive words as positive (notated by 1) and with more negative words as negative (notated by 0).

Apply the sentiment analysis method to the manually tagged test data set, then the calculated error rate is 21.6%, which is much better than our expectation.

Combination: Eventually, combine the processed data sets including evaluated sentiment with the **Regional Cases** data set, to obtain, for each tweet, the corresponding American state-level local COVID data, in the same state and on the same day.

Model

Expectation: When the COVID condition are more severe, to be more specific, higher amount or increment of death o infection, or higher rate of death, the attitude of local people tend to be more negative.

Model 1: Logistic Regression

Since the calculated sentiment outcome is 0-1 variable, then we consider the simple logistic regressions with the sentiment as response and with cumulative number of infection, increment of infection, cumulative number of death, increment of death, rate of death respectively as the predictor. Since the five predictors tends to be highly relative, in case of multi-collinearity, we decide not to add all these predictors into one model.

Model Assumption:

- **Independent**: all the calculated tweet sentiment are independent, or no interaction between tweets.
- **Linearity**: the predictor and intercept linearly impact the response.
- **Link**: take the `logit` function as the link function.
- **Distribution**: the responses follow independent binary distribution.

Take the predictor log-transformed death rate for instance.

In the fitted logistic regression model, although, the sign of the coefficient of the predictor meets our expectation, the coefficient is not statistically significant with large p-value, and the difference between the residual deviance and null deviance is smaller the threshold 1.

And in other simple logistic regression models with other predictors, none is statistically significant and with expected deviance decrease.

Model 2: Linear Regression Model with Transformed Data

Group the data set by the states, and the calculate the mean of the evaluated sentiment outcome of each state and the mean of the death rate of each state. Apply the `logit` function to both the mean of sentiment, ranged from 0 to 1, and the mean of death rate, draw the scatter plot to inspect the trend.

Model Assumption:

1. **Independent**: all the calculated tweet sentiment are independent, or no interaction between tweets.
2. **Linearity**: the predictor and intercept linearly impact the response.
3. **Link**: take the identical function as the link function.
4. **Distribution**: the responses follow normal distribution.

Fit the linear regression model with the transformed sentiment mean.

The sign of coefficient of the predictor coincides our expectation. However, as the former simple logistic regression models, the model is not statistically significant, nor the coefficient of the predictor.

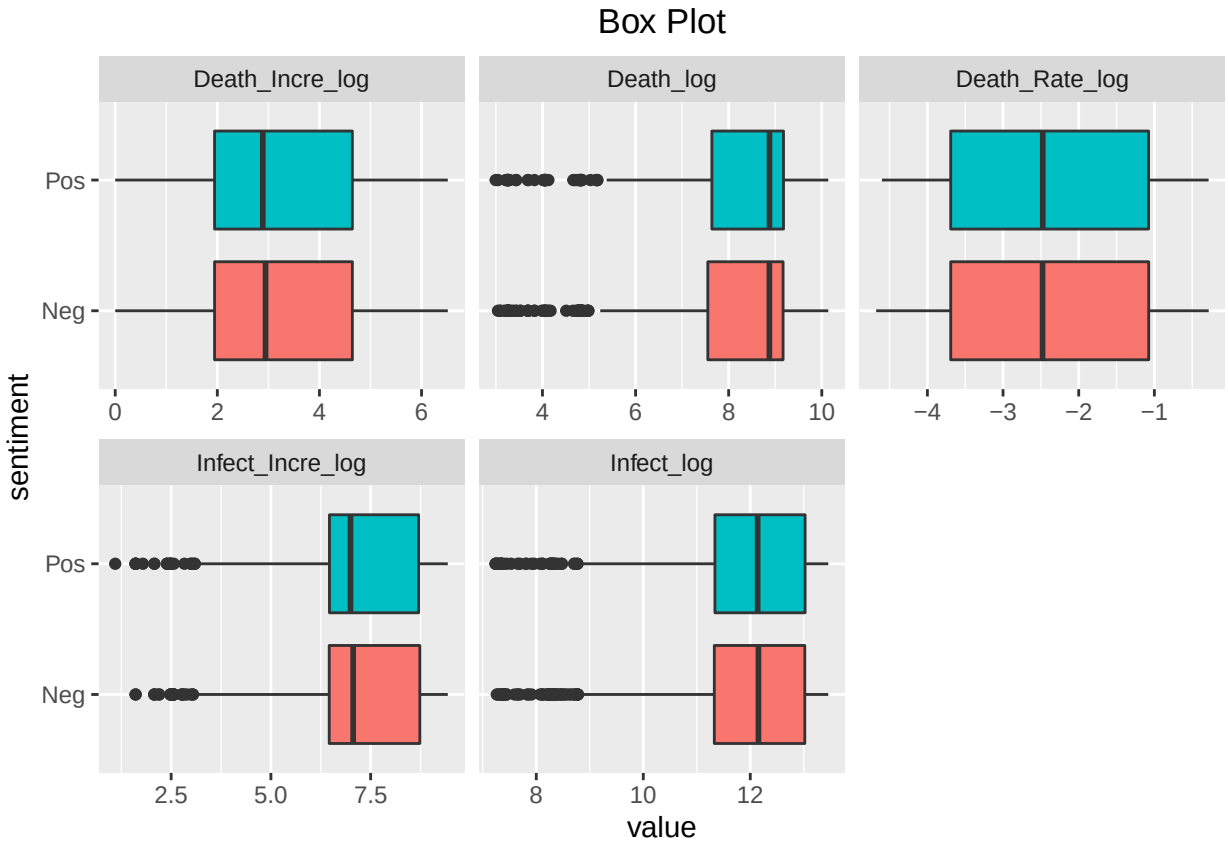
Model Interpretation and Discussion

The logistic regression and linear regression models do not present enough statistical significance. Therefore, at this point we do not have enough evidence to answer our statistical question, nor the broader question

at stake. At this moment, we can neither prove that the COVID infections and deaths numbers influence people's sentiments on tweets nor disprove it. Regarding the lack of statistical power in our models, we provide the following possible explanations.

To begin with, the problem might come from the dataset. The dataset we are using includes tweets gathered worldwide within a 1-month period. Since for each tweet, we use the COVID data based on the date and location it was released, such a short time period leads to overlapping data. Specifically, many tweets that are posted on the same day and same location(Country specific) may have opposite sentiment, thus generating noise that mislead the models. Further more, the dataset consists only of tweets with hashtag `#covid19`. Whether such a dataset is indeed a good representation of all the tweets posted during the pandemic is yet unclear. To address the issues above regarding the dataset, we plan to take the following approaches. To remove the possible bias in the dataset, we plan to generate a new dataset randomly sampled from all the tweets posted during the pandemic, with preferably larger range of dates. To reduce the noise in the dataset, we plan to generate a daily sentiment label for each day in the dataset, by calculating the majority sentiment(good vs. bad) from all the labeled tweets in each location on each date.

Apart from potential issues with the dataset, the models could carry potential problems. In the sentiment analysis model where the sentiment labels are generated, the true accuracy of the model is unclear. The current model calculates the sentiment score for each tweets based on a predefined lexicon in R, which is essentially a dictionary that maps each word to an either positive or negative score. In testing, we received roughly 80% accuracy when predicting on another labeled tweets dataset. However, the model's true ability to perceive the sentiment is unclear. Since sentiment is quite subjective and different people might agree on opposite labels for the same tweet. The logistic regression and linear regression models(Fit on the COVID trend variables and the sentiment labels) might also under-perform in this case since the relation between the COVID trends and the tweets sentiment might not be linear, if it exists at all. To address potential issues with the model, we plan to compare multiple sentiment methods and their application in the context of the problem. Furthermore, we plan to include more models in fitting the COVID trend variables and the sentiment labels. Polynomial regressions and models such as Random Forest are currently on our to-do list. We wish to explore the strength for each model and compare their applicability in terms of both goodness-of-fit and interpretability. In the end, we expect to compare all applicable models and choose the best one. If multiple models turn out to be best fits, we will compare the conclusion from each model and perform model specific interpretation and inference on the results. Based on the results, we will then establish an answer to our statistical question as well as the broad question.

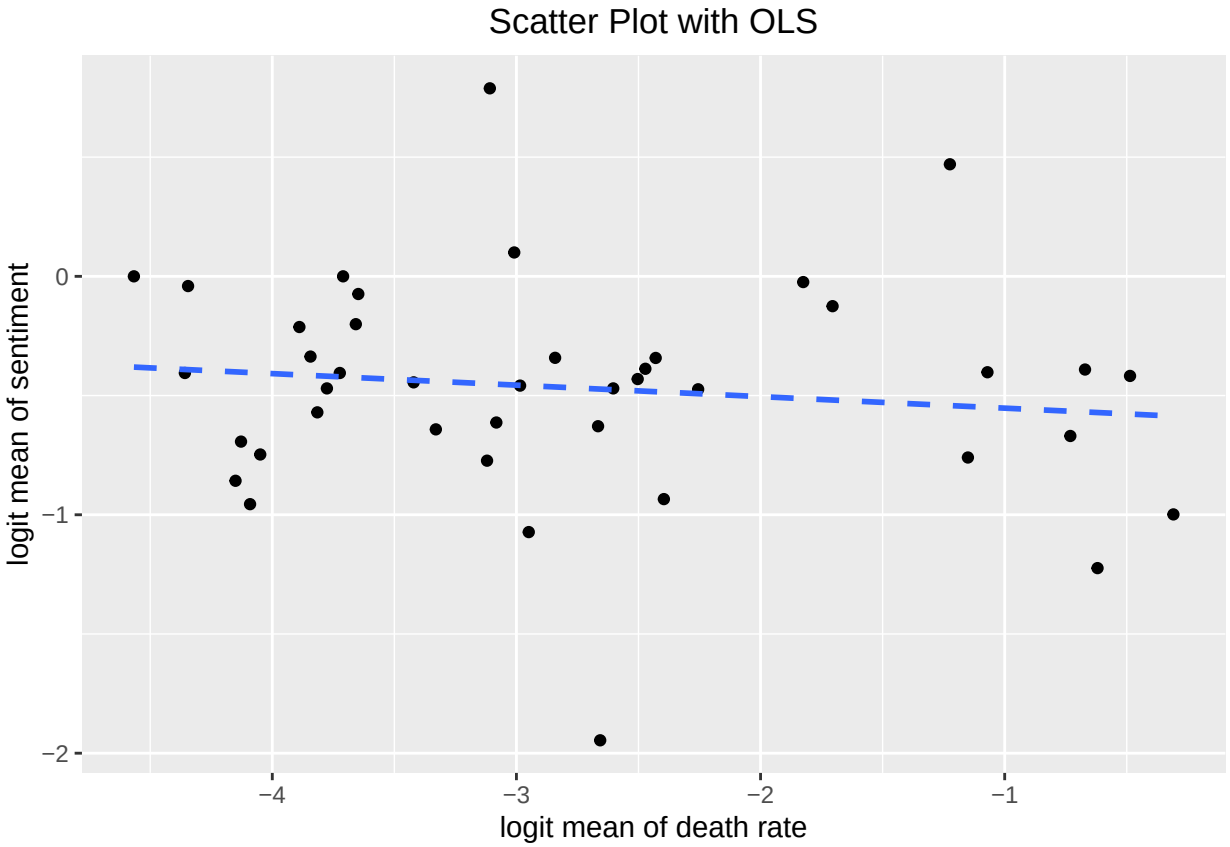



```
## `summarise()` ungrouping output (override with `.groups` argument)

## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 2 rows containing non-finite values (stat_smooth).

## Warning: Removed 2 rows containing missing values (geom_point).
```


Scatter Plot with OLS
logit mean of sentiment
0
−1
−2
−4
−3
−2
−1
logit mean of death rate