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

We have three datasets: Tweets, Sentiment and Covid Cases by Region.

Tweets Datasest: This dataset contains tweets from July 25, 2020 to August 30, 2020 with #covid19 hashtag. The column names include usernames, locations(countries) and texts etc.. Although there are 181061 tweets, lots of the them are not written in English. Besides, some of the countries only have small number of observations. Based on the information above, we decided to only focus on the tweets in the US. We constructed the response variable(sentiment of tweets) by performing a sentiment analysis to the "texts" column.

The major limitations of this dataset include:

- The lack of time span - the data is ranged within a month. Such limitation would result in lack of variation in our dataset, so it would be difficult to find the COVID-19 cases and deaths' impact on people's emotion.
- This dataset only collected tweets with #covid19 hashtag. As we discussed above, some users might not add this hashtag when they post tweets. In fact, lots of the tweets with #covid19 hashtag are news reports without any personal opinion. Hence, such limitation might lead to inaccurate prediction of the relationship between COVID-19 cases and deaths' and people's emotion.
- All the tweets are truncated to 140 characters and most of them are ended with apostrophe, which might lead to misunderstandings and risk the accuracy of the sentiment analysis.

Sentiment Dataset: This dataset contains approximately 40,000 tweets. All of the tweets are tagged with 5 sentiment levels: "Extremely Negative", "Negative", "Neutral", "Positive" and "Extremely Positive".

Covid Cases by Region Dataset: This dataset collects information from 50 US states, the District of Columbia, and 5 other US territories. The dataset provides positive and negative test results, pending tests, as well as total hospitalizations, deaths and recovered population. We combine this dataset with the Tweets dataset to analyze whether variables such as the death rate, number of the confirmed cases, rate of the deaths and recovered are associated with the users' emotion.

Pre-Processing

Location Cleaning: Since the location information in Tweets dataset does not have a fixed format and it has some typo issues, it is necessary to conduct text cleaning. The objective of location cleaning is to transform the location column from upper cases to lower cases, search for location strings containing the special substring (e.g., "usa"), search for state name and city name, and extract or convert the results to the corresponding state name.

Text Cleaning: We cleaned the tweets text by reomving punctuations, non-alphanumeric symbols, links that start with "http" or "https" etc. After the removal of the unwanted ingredients, the remaining text was inputed into the sentiment analysis and labelling.

Sentiment Analysis: We conducted the sentiment analysis through the 'bing' lexicon(a built-in lexicon in R). The main idea was to evaluate the sentiment (score) of each word based on the lexicon and use the sum of score as the sentiment of the whole text. We omitted the tweets with equal amount of positive and negative words because these neutral tweets might be less informative. We labeled the tweets with more positive words as positive (notated by 1) and tweets with more negative words as negative (notated by 0).
Assessment: Next, we mapped the 5-level sentiment of the manually tagged tweets data set to 2-level label "Positive"(1) or "Negative"(0). Such dataset is treated as a test dataset to validate the sentiment analysis method. We applied the sentiment analysis method above to the dataset and found that the calculated error rate was 21.4%, which was much better than our expectation.

Combination: Finally, we combined the results of the sentiment analysis and the Regional Cases dataset to a dataset that contained the sentiment score of each tweet and their corresponding state-level COVID-19 data(number of cases/deaths).

Model

H_0 : The change of covid19 pandemic condition, including the change of new cases and new deaths, does not influence the sentiment of tweets.
H_1 : The change of covid19 pandemic condition, including the change of new cases and new deaths, influences the sentiment of tweets. In particular, as confirmed cases and deaths rise, the likelihood of tweets sentiment reaches 0(negative) rises.

Model 1: Logistic Regression

We check the respective box plots of potential predictors grouped by sentiment, which is in the attachment: Figure 1, and there seems to be no difference in the potential predictors between each sentiment group. Since the calculated sentiment outcome is 0-1 variable, 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 0.57, and the difference between the residual deviance and null deviance is smaller the threshold 1. Then, we calculated the McFadden's pseudo r-squared of the model, which is around 8.64×10^{-5} and excessively small, collaborating that the explanatory ability of the model is extremely weak. The main output is in the attachment.

And in other simple logistic regression models with other predictors, none predictor is statistically significant with expected deviance decrease, and none model holds strong explanatory ability with large pseudo r-squared.

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, which is the in the attachment: Figure 2.

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 mean sentiment and transformed mean death rate.

The sign of coefficient of the predictor coincides our expectation. However, as the former simple logistic regression models, the model with trivial r-squared 0.0054 is not statistically significant, nor the coefficient of the predictor with large p-value 0.65. The main output is in the Attachment.

Model Interpretation and Discussion

The logistic regression and linear regression models do not present enough quality. The goodness-of-fit is not good enough. 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. 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.

Attachment

Plot

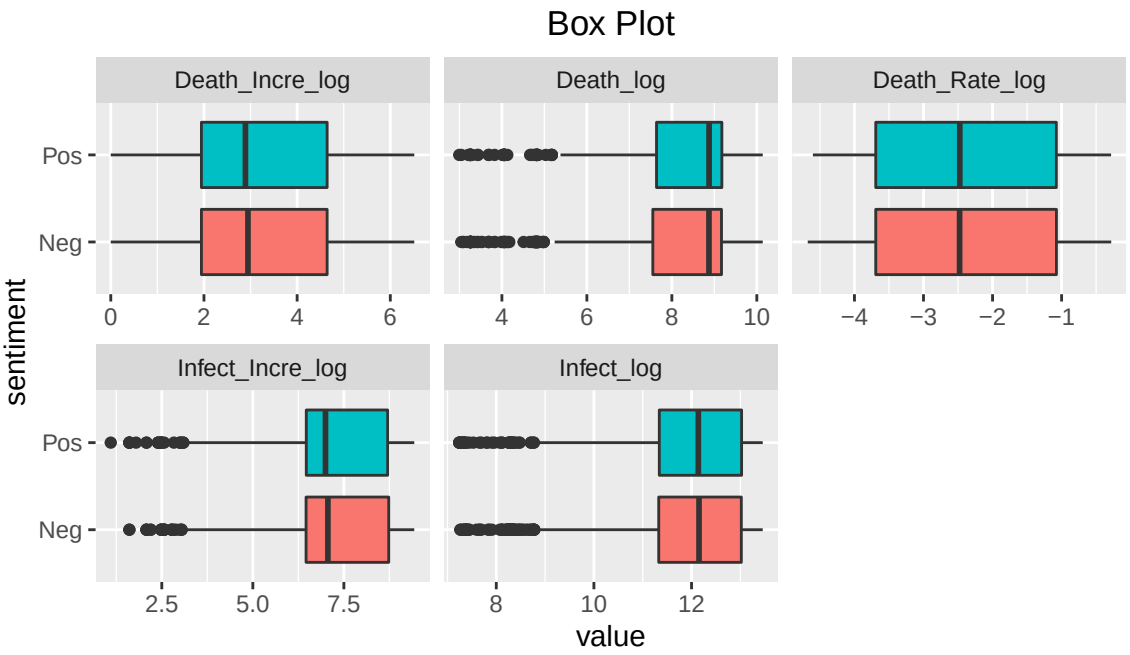

Figure 1: Box Plot of Sentiment and 5 Predicotrs Respectively

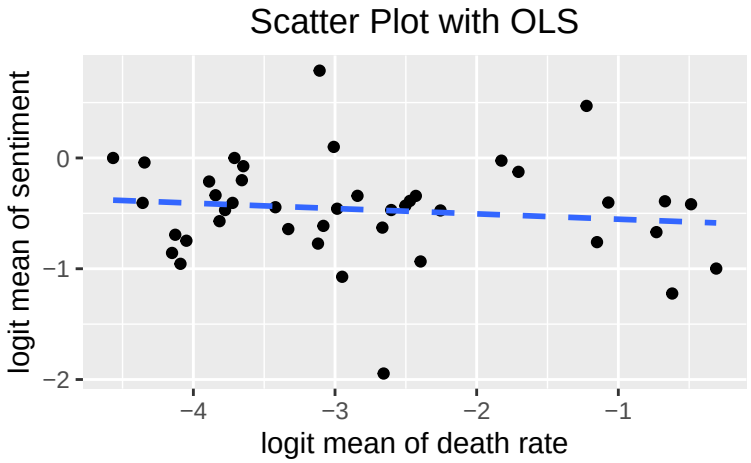

Figure 2: Scatter Plot of Logit Mean Sentiment and Logit Mean Death Rate

Model Output

Model 1

```
model.glm.death_rate <- glm(data = data.senti_covid %>%
                              filter(!is.na(recovered)) %>%
                              mutate(death_ratio_log = log(death/recovered)),
                            formula = sentiment ~ death_ratio_log,
                            family = binomial())
summary(model.glm.death_rate)
```

```
##
## Call:
## glm(formula = sentiment ~ death_ratio_log, family = binomial(),
##     data = data.senti_covid %>% filter(!is.na(recovered)) %>%
##         mutate(death_ratio_log = log(death/recovered)))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0162  -1.0033  -0.9905   1.3606   1.3830
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.47638    0.08613  -5.531 3.18e-08 ***
## death_ratio_log -0.01809    0.03187  -0.568     0.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3730.7  on 2782  degrees of freedom
## Residual deviance: 3730.4  on 2781  degrees of freedom
## AIC: 3734.4
##
## Number of Fisher Scoring iterations: 4
```

```
pscl::pR2(model.glm.death_rate)[4]
```

```
## fitting null model for pseudo-r2
```

```
##     McFadden
## 8.638175e-05
```

Model 2

```
model.lm.death_rate.states <- data.senti_covid %>%
  filter(!is.na(recovered)) %>%
  group_by(state) %>%
  summarise(death_rate_mean_log = log(mean(death/recovered)),
            sentiment_mean_logit = faraway::logit(mean(sentiment))) %>%
  ungroup() %>%
  filter(sentiment_mean_logit != 0) %>%
  lm(formula = sentiment_mean_logit ~ death_rate_mean_log)
```

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

```
summary(model.lm.death_rate.states)
```

```
##
## Call:
## lm(formula = sentiment_mean_logit ~ death_rate_mean_log, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45438 -0.23020  0.02719  0.15587  1.26727
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.56628    0.18351  -3.086  0.00378 **
## death_rate_mean_log -0.02814    0.06178  -0.455  0.65139
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4597 on 38 degrees of freedom
## Multiple R-squared:  0.005429,   Adjusted R-squared:  -0.02074
## F-statistic: 0.2074 on 1 and 38 DF,  p-value: 0.6514
```