Final Report

1. Introduction

1.1 Background

COVID-19 pandemic has been one of the most severe issues of the world as it appears to be equivalent to a world war. There are tons of adverse ramifications physically and mentally affected by this pandemic. Obviously, it leads to serious illnesses and death without any effective mitigation approaches at the beginning. In another aspect, COVID-19 has been emerging to destroy world growth in economics which drives all activities and developments to the world. This interests the authors to come up with the question of "How does COVID-19 affect the growth of world economics?".

One of the most apparent indicators of world economics is the unemployment rate which represents whether the world is moving forward effectively, and in turn people have the ability to pay for their own livings. High unemployment rate refers to the problem that companies have less capability in taking care of their employees leading in both situations of unemployment and laying off while it slows down the pace of world growth as it should be. According to the Congressional Research Service (CRS) research, it is surprising that the highest unemployment rate occurred in the era of COVID-19 [1], much higher than any pandemic in the history of the U.S. This can be initially confirmed that there are some significant relationships between the COVID-19 pandemic, which allows us to perform data analysis regarding establishing the model to explain the variables and confirming that the relationships exist.

1.2 Objective

- To refine a broad question to a more specific question that can be answered by data analysis.
- To perform an exploratory data analysis to examine COVID-19 variables and unemployment rate data distribution and relationships.
- To construct the statistical model to describe the variables and relationships as well as to perform hypothesis testing to confirm the relationships.
- To interpret the results and answer the questions of interest.

2. Methodologies

2.1 Refining the question

According to the broad question, "How does COVID-19 affect the growth of world economics?". Our first objective is to refine this question into a more specific question which possesses the characteristics of a good question. After a thorough discussion, the refined question is

"Are there any significant relationships between COVID-19 figures and the unemployment rate in the U.S.?"

This is an exploratory question which requires the use of statistical modeling to describe the data and variables. Furthermore, this can be an inferential question since we aim to test for the significance of the relationship between these variables. The following criteria also explain why the authors decide to select this refined question.

- This question is of interest to individuals as well as the organization in every industry in the U.S. since people these days have been paying attention to this large-scale pandemic. Besides, the research from CRS indicates that the unemployment rates sharply increased to the peak in every sector during this period. Therefore, results analyzed in this topic could help prevent incoming loss in a timely manner.

- It is established in the plausible framework since it is grounded by the current situation of COVID-19 supported by news and numerous figures regarding unemployment from various sources. It follows that this question can be answered by statistical approach.

- As we focus on one specific criterion of world economic growth which is an unemployment rate, this question is scoped down and possesses the specificity.

2.2 Design of experiments

2.2.1. Data collection phase

According to the study, we are searching for data related to COVID-19 and some necessary economic factor to be our predictors while the response variable is an unemployment rate in the U.S. We collect an unemployment rate data each month in every state in the U.S. from www.kff.org (Kaiser Family Foundation). We obtain the COVID-19 figures data such as number of new cases, death in the U.S. from www.cdc.gov and vaccination rates from ourworldindata.org as well as the lockdown days data that is collected from www.usatoday.com. We also consider including some economic factors that might relate to the unemployment such as average GDP and personal income in our analysis. The data is collected each quarter and state from www.bea.gov/data/. All the data are selected during January 2020 to August 2021 and merged into a well-structured dataset classified by states and quarter. The sample dataset is shown in Appendix.

2.2.2. Exploratory data analysis

From the question we set, we investigate our data set and variables to not only check whether the data has any problems, but also to explore some interesting insights regarding the variables that might be useful for the next steps of analysis. Therefore, we select some the interesting visualization figures with insights as follows: (All results are shown in Appendix)

- (Figure 1) The minimum unemployment rate for Southwest region is relatively higher than other regions although the histogram (Figure 2) indicates that there are small number of states compared to others
- (Figure 3) The distribution of unemployment rate is positively skewed demonstrating that the average of unemployment rate (5% to 10%) is smaller than the median.

- (Figure 4) The scatter plots of unemployment rate versus COVID-19 factors shows that there might be a chance that the several predictors are linearly related to the response variable. However, we can make sure our assumptions by testing them statistically in the next section.

Figure 1

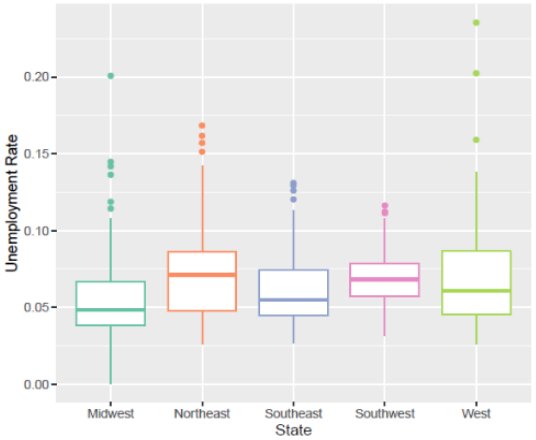


Figure 2

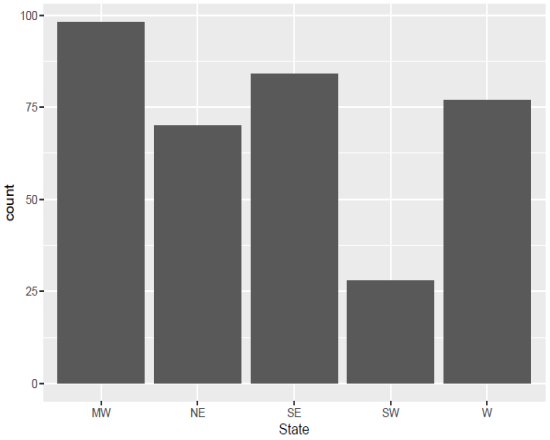


Figure 3

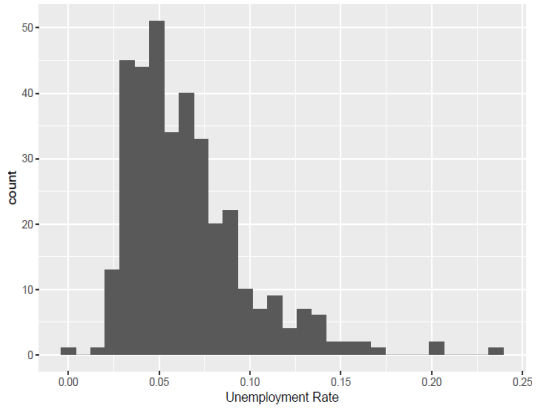
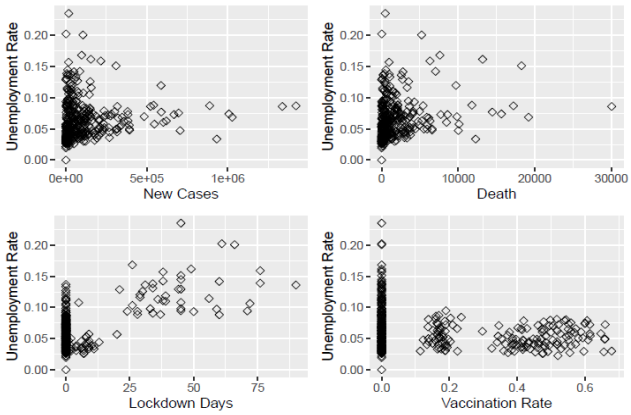


Figure 4



2.2.3. Statistical modeling

According to the question we set, we consider applying statistical modeling techniques to describe the variables and relationships. Our initial model starts with:

$$\boldsymbol{Unemployment\ rate} = \boldsymbol{\beta_0} + (\boldsymbol{nCase})\boldsymbol{\beta_1} + (\boldsymbol{nDeath})\boldsymbol{\beta_2} + (\boldsymbol{vaccRate})\boldsymbol{\beta_3} + (\boldsymbol{lockdownDay})\boldsymbol{\beta_4} + (\boldsymbol{state})\boldsymbol{\beta_5} + (\boldsymbol{quarter})\boldsymbol{\beta_6} + (\boldsymbol{GDP})\boldsymbol{\beta_7} + (\boldsymbol{personalIncome})\boldsymbol{\beta_8}$$

After fitting the model using *lm* function to estimate the coefficient, it appears that the model has too many variables and coefficients since states are separately estimated which are difficult to interpret as shown in Appendix. This brings us to group the states variable into different regions, “Northeast”, “Southeast”, “Southwest” and “West”. Then we continue to fit the model again and the result is shown as follows:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.583e-02  3.133e-03   8.243 3.78e-15 ***
stateNortheast   1.093e-02  3.075e-03   3.554 0.000434 ***
stateSoutheast   3.642e-03  2.749e-03   1.325 0.186104
stateSouthwest   9.843e-03  3.969e-03   2.480 0.013620 *
stateWest        1.079e-02  2.809e-03   3.840 0.000147 ***
quarter20-q2     6.358e-02  5.342e-03  11.902  < 2e-16 ***
quarter20-q3     4.432e-02  3.692e-03  12.006  < 2e-16 ***
quarter20-q4     2.642e-02  4.232e-03   6.243 1.29e-09 ***
quarter21-q1     1.736e-02  5.775e-03   3.005 0.002852 **
quarter21-q2     8.815e-03  1.142e-02   0.772 0.440905
quarter21-q3     4.330e-03  1.256e-02   0.345 0.730451
new.cases       -1.283e-08  1.388e-08  -0.925 0.355807
death            8.349e-07  6.885e-07   1.213 0.226115
lockdown.days    5.316e-04  1.167e-04   4.557 7.28e-06 ***
GDP.USD.         2.883e-09  3.148e-09   0.916 0.360445
Personal.Income  6.353e-09  4.612e-09   1.378 0.169266
vaccination.rate 1.791e-02  2.342e-02   0.765 0.445051
```

According to the result, it shows that only “lockdown.days” factor is significantly linearly related to unemployment rate. As for other COVID figures such as “new.cases”, “death” and “vaccination.rate”, as well as economic variables, do not have any significant relationships to unemployment, whereas quarters and regions appear to be significantly related to our response variable.

We also build another model of logistic regression to evaluate the probability or rate of unemployment based on these predictors. The result is shown below:

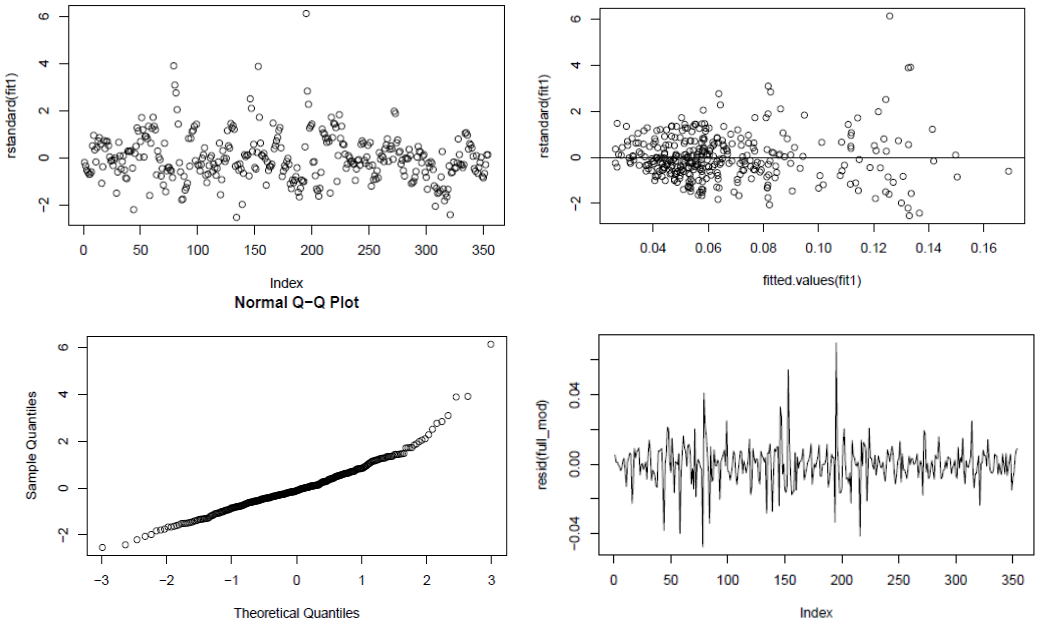
```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.423e+00  8.518e-01  -4.018 5.87e-05 ***
stateNortheast   1.867e-01  6.844e-01   0.273    0.785
stateSoutheast   6.960e-02  6.386e-01   0.109    0.913
stateSouthwest   1.748e-01  8.772e-01   0.199    0.842
stateWest        1.821e-01  6.388e-01   0.285    0.776
quarter20-q2     1.077e+00  1.126e+00   0.956    0.339
quarter20-q3     8.163e-01  9.150e-01   0.892    0.372
quarter20-q4     5.330e-01  1.064e+00   0.501    0.616
quarter21-q1     3.539e-01  1.454e+00   0.243    0.808
quarter21-q2     1.473e-01  2.854e+00   0.052    0.959
quarter21-q3     4.127e-02  3.131e+00   0.013    0.989
new.cases       -3.944e-08  2.744e-06  -0.014    0.989
death            3.993e-06  1.256e-04   0.032    0.975
lockdown.days    4.812e-03  2.024e-02   0.238    0.812
GDP.USD.         5.350e-08  6.672e-07   0.080    0.936
Personal.Income  7.435e-08  9.478e-07   0.078    0.937
vaccination.rate 4.735e-01  5.735e+00   0.083    0.934
```

It appears that the logistic regression model cannot capture any relationships between any predictors and unemployment rate. Ultimately, combined with the result obtained from previous linear model, we are not able to confirm that there are significant relationships between COVID-19 factors and unemployment rate.

2.2.4. Model validation

I) Linear regression model validation

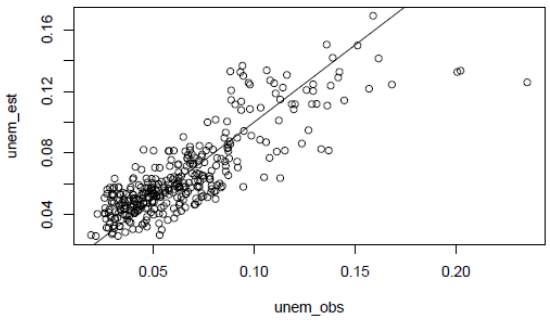
Even though there is some obscure relationship between these variables, we also perform model validation test to see how the linear models perform. The results are demonstrated below:



It can clearly be seen that all the assumptions of linear modeling are violated. The first two plots testing the assumption of constant variance and linearity respectively show that errors are not random which means the variance is not constant and the parameters do not linearly interact. Moreover, Q-Q plot also shows some strange alignment of plots at each tail indicating that the observations are not normally distributed. Lastly, for the plot of independence test, although it is not clear whether there is a pattern in the plot, we perform Durbin-Watson test and obtain that the p-value is large. In other words, we reject the null hypothesis that there is no correlation between observations. (The test is shown in Appendix)

II) Logistic regression model validation

As for binomial model, we also test for model validation as we plot the observed unemployment rate versus the estimated unemployment probability. It shows some strange behavior for a relatively high rate of unemployment. We also include the model using "quasibinomial" in Appendix section to check for the model dispersion and we obtain the dispersion parameter of 0.005 which means we should consider other model that fits the data.



3. Analysis interpretation and discussion

In this section, we would like to succinctly summarize what we have obtained and learned so far. There are several key takeaways that we would like to point out as follows:

- Based on the visualization, we can see that the scatterplots from figure 4 have a huge amount of data points at zero for each predictor. This might be the reason that the linear model does not fit this type of data. Also, the scatterplots for new cases and death have a similar pattern. We might consider modeling the data either by adding interaction term or using non-linear approaches. In this case, the hierarchical model is also interesting since we have a grouped level of states (or region).
- As for the modeling part, it might be possible that only COVID-19 factors cannot capture all the variability of the response, so that we might consider including other related variables. In a meantime, all the COVID-19 figures are based on the number of populations which means, for vaccination rate, cannot represent clearly on how large it is since different states / regions have totally different population. Therefore, it is recommended to have populations figure as one of the columns.

4. Answer to the refined question

According to all the results obtained, it tends to be inconclusive that there are significant relationships between the COVID-19 factors and unemployment rate since the p-value for each variable in the model summary is relatively larger than 0.05, the typical significance level. However, this might be because the linear model is not the best fit for describing the variables, so that it might be better if we model the data in non-linear modeling approach.

Additionally, another suggestion is that there might be other COVID-19 factors that relate to unemployment better such as medical expenses for COVID patients, committed suicide rate due to COVID-19 anxiety or types of business that severely affected by COVID etc. The variables we used in this analysis are only a representation of COVID-19 figures, so that we cannot clearly state whether COVID-19 factors do not correspond to unemployment.

5. Answer to the broad question

Since we are not able to conclude the answer to the refined question, for our bigger question which refers to the global economic growth, it also cannot be clearly concluded that COVID-19 has a significant effect to the growth of world economics. Similarly, we can change the response variable representing economic growth rather than unemployment rate such as world-leading company's stock price or the Consumer Price Index (CPI) or Gross Domestic Product (GDP) etc. Another aspect is that it can be expanded in analyzing this topic in such a way of studying the world economic factors across the globe.

Appendix

Data sources and References

1.1 Data used in the analysis

- www.kff.org
- www.cdc.gov
- ourworldindata.org

1.2 Source of research and information

[1] Gene Falk, Romero P., Nicchitta I., Nyhof E. Unemployment rates during the COVID-19 pandemic. The Congressional Research Service, U.S.A.