STAT 601 Final Project Report

Section 002 Group 3

Mengqi Xu
mxu276@wisc.edu

Ouyang Xu
oxu2@wisc.edu

Shubo Lin
slin268@wisc.edu

Zhenpeng Shi
zshi224@wisc.edu

Date: December 15, 2020

Abstract

We applied a multiple linear regression model to measure the impact of lockdown on the spread of COVID-19. We break down lockdown into several factors and collect data from google mobility report and google trends. All the codes and data are available on https://github.com/ShermanSteinke/STAT601/tree/master

1 Question

When COVID-19 was first spotted in China, it was considered as merely a kind of mutated flu. A small outbreak in Wuhan led to a quick response from the Chinese government and then from the world. Despite the quick lockdown done by the many countries, the COVID-19 has become a global pandemic since its first outbreak at the end of 2019. Now the outbreak does not show any sign to stop or even slow down. Many countries and states all over the world are facing a second giant outbreak as the winter approaches. As a result, many countries in Europe and states in the United States have declared a second lockdown on their cities. The first massive lockdown earlier this year has led to two completely different opinions towards it. A lot of people believe that locking a city down reduces communication and the chance of virus transmission. On the other hand, quite a few people proposed that a complete lockdown violates human rights and makes many people unable to work which may worsen the pandemic. Moreover, businesses will shut down due to massive lockdown, and then the economy collapsed. The intense arguments on the legitimacy of massive lockdown trigger our interest in studying the impact of lockdowns, to be more specific, to study how effective the lockdowns were and what factors that impact the effectiveness of the lockdowns were. Then finally we will conclude which factor(s) play important roles in inhibiting the spread of COVID-19.

2 Data

From the perspective of a statistician, the main goal of the project is to quantify the lockdown and the pandemic. Here we do not simply assume lockdown is a single independent binary variable, but we believe that the factor, lockdown, can be disintegrated. To be more concrete, we do not treat lockdown as a single binary factor, which is either 0 or 1, but a factor may consist of many other numerical or categorical factors. One of our biggest challenges of our project is to make the lockdown measurable. We may find a direct projection of the lockdown by measuring other aspects of the pandemic. In other words, we need to find something measurable that can directly reflect the level of lockdowns. As we work through the topic, we have found that the topic is not only a statistical problem but also relates to Economics, Geography, and Politics, etc. We will try to decorrelate the irrelevant factors and build a statistical correlation between each major factor.

The COVID-19 dataset of the United States was obtained from the COVID-19 Tracking Program, and that of Italy is from Dati COVID-19 Italia collected by *Sito del Dipartimento della Protezione Civile - Emergenza Coronavirus: la risposta nazionale*. Data for total cases, daily new cases, total tests, daily new tests, etc., were extracted for all 50 states from March 20, 2020 to Nov 20, 2020. Lockdown timeline is obtained from the Wikipedia COVID-19 pandemic lockdown. This dataset is very concrete for case tracking.

We've tried to divide the lockdown into several different ways, such as the restriction of retail and recreation, grocery and pharmacy, parks, transit stations, workplaces, and so on depending on Google COVID-19 Community

Mobility Reports. We collect data in all the states from March 20, 2020 to Nov 20, 2020 which is open to the public by Google location services, because the length of stay in the above places is a good measurement tool that may partially provide some information on lockdown. Users may provide Google with different types of location information, and we can infer how the length of stay at different places change compared to a baseline, which is the median value, for the corresponding day of the week, during the 5 weeks from Jan 3 to Feb 6, 2020. Moreover, trends data is obtained from Google trends. Trends, or google search index, shows the search volume of different queries over time. Notice that the trends data are not absolute values but relative values after shifting and scaling into a range 0-100. It can help illustrate people's awareness of a certain term during a certain time. In this research , we use the term 'face mask', 'hand sanitizer', 'quarantine' and 'COVID-19 testing'.

3 Model

3.1 Model 1: Binary Lockdowns

3.1.1 *Model setup*

Initially, we consider the Multiple Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 i + \beta_2 C_i + \epsilon, \qquad i = 1, 2, \ldots, n$$

where

- Y_i denotes the cumulative cases on the i^{th} day;
- i denotes the i^{th} day after the first day we counted;
- C_i denotes the lockdown status; $C_i = \begin{cases} 0 & \text{if state was not locked down during the selected period,} \\ 1 & \text{if state was locked down during the selected period,} \end{cases}$
- β_0, β_1, and β_2 denote regression coefficients corresponding to continuous date and binary lockdown variable.
- $\epsilon_i \sim N(0,\ \sigma^2)$ denotes the i^{th} random error term.

3.1.2 *Interpretation*

The model above is just a rough prototype of what we expect to do, which only considers two factors: date and lockdown. Briefly speaking, we suppose the lockdowns could decrease the daily new confirmed cases by setting a very simple model saying that there is only one factor, the lockdown, that truly matters, and we set this value binary, which is either 0 or 1. If the linearity holds, then we should be able to find out the impact of lockdowns on different states by looking at the coefficient of C_i. If the lockdowns played a positive role in preventing the spread of COVID-19, β_2 should be negative and the more effective lockdown is, the larger magnitude of β_2 should be, and vice versa.

3.1.3 *Result*

Based on the first model, our outcome on states of the United States which had officially declared statewide lockdowns are in Appendix 1: Table 1

We have the result above by modelling each state separately, because the data we have got is not only sorted by states but also depends on multiple local factors. We cannot draw a nationwide conclusion by summing the numbers and fit the data to the entire country because in many cases, the lockdown situation were completely different in different states and the policies of lockdowns from each local government were different as well. We have applied the multiple linear models to the data from several states. The coefficient of C_i, β_2, is expected to be negative according to our background setting. However, after studying the cases of Michigan, California, New York, Illinois and Connecticut, we found that not all states followed the assumption. The models built on California, Michigan and New York have a positive coefficient on C_i, which indicates lockdowns had a negative effect on controlling the spread of COVID-19. After taking into consideration the actual situation of each state at the time, we conclude that due to heavy infections in these states, the testing capability of local health authorities might not be able to cover all the infected; therefore, the reduced number of newly confirmed might not be reflected on the data. On the other hand, the models on Illinois and Connecticut worked very well on reflecting the impact of lockdowns. The coefficients of C_i are not only negative but also very close which implies these two states were facing some similar situations.

3.2 Model 2: Multiple Mobility Linear Regression on Lockdowns

3.2.1 *Model setup*

Then, after considering the actual situation and fitness of the old model, we have modified our model into:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon, \qquad \epsilon \sim \mathcal{MVN}(\boldsymbol{0}_{n\times 1},\ \sigma^2 \boldsymbol{I}_{n\times n})$$

where

- $\boldsymbol{Y} = [Y_1\ Y_2\ \dots\ Y_n]^T$ denotes the response variable matrix. Initially, we set the response variable to the new positive cases by days in the specified state or region.
- $\boldsymbol{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,5} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,5} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,5} \end{bmatrix}$ denotes explanatory variable matrix with $X_{i,1}, X_{i,2}, \dots, X_{i,5}$ corresponding to the mobility of "retail and recreation", "grocery and pharmacy", "parks", "transit stations" and "workplaces", respectively;
- $\beta = [\beta_0\ \beta_1\ \dots\ \beta_5]^T$ denotes the intercept and coefficients of the explanatory variable.

3.2.2 *Model modifications*

First of all, in order to guarantee the model accuracy, all assumptions need to be satisfied by all means. According to the definition of linear regression, we should have the following assumptions:

- Linearity: There is a straight line relationship among response variable $\boldsymbol{Y}$ and explanatory variables $\boldsymbol{X}$.
- Independence: Response variable Y_i are independent.
- Equal Variance: $\text{Var}(\boldsymbol{Y}) = \text{Var}(\epsilon) = \sigma^2 \boldsymbol{I}_{n\times n}$
- Distribution: $Y_i \sim N(\boldsymbol{X}\beta,\ \sigma^2 \boldsymbol{I}_{n\times n})$
- Low level of multicollinearity: explanatory variables have very low correlations with each other.

However, due to great complexity of the real situation of different states, many assumptions failed to hold based on the original data. In order to implement the proper assumptions, we have done several modifications to the model. First of all, the initial model 2 has very high multicollinearity by calculating the VIFs (variance inflation factor) of each explanatory variable. We noticed that the transit stations have very high VIF compared with other variables, which means mobility of transit stations is highly correlated with mobility of other areas. We considered it reasonable because if the mobility of other areas has changed, then it is very likely for people to move to transit stations and thus the mobility of transit stations will change. Therefore, we removed the transits station variable because it can be reflected by other variables due to high multicollinearity,and the VIFs can be checked in Table.5, which is smaller than before. Then we applied Box-Cox transformation(Box and Cox (1964),Li (2005)) to the response variable Y_i in order to remove heteroscedasticity and increase normality of the data:

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln Y_i & \text{if } \lambda = 0, \end{cases}$$

where

- Y_i denotes new positive cases matrix by days in the specified region

- The parameter λ is estimated by maximizing the profile likelihood function

After the transformation, all assumptions hold for most of the states in the US and regions in Italy. Moreover, the impact of lockdowns was not instant, which means the impact needs days to be reflected in the number of new cases. Thus, we set a lag duration which counts the average positive increase between 7 days and 14 days after the number of new cases was counted according to research on the average time for patients to show symptoms. We finally apply the modified data to the model above and generated a reasonable result.

Moreover, we noticed that daily new cases vary a lot according to daily test capacity and it might be undercounted. To add more robustness to our model, we tried to change our response variable to a weighted sum between average daily new cases from data and an average theoretical daily new cases computed from death data:

$$Y_i = w_1 \times \frac{1}{8} \sum_{j=i+7}^{i+14} P_j + w_2 \times \frac{\frac{1}{15} \sum_{j=i+14}^{i+28} D_j}{DeathRate}$$

where

- Y_i denotes the i^{th} response variable.
- w_1 and w_2 denote the weight we set to calculate weighted daily new positive cases. In this research, we set $w_1 = w_2 = 0.5$ but it needs further adjustment to get more accurate data. However, the reference was very difficult to find, thus such transformation might become less accurate.
- P_j denotes the positive increase on j^{th} day.
- D_j denotes newly reported deaths on j^{th} day.

[Time lag 7-14 days and 14-28 days are based on the incubation period of Coronavirus disease(Lauer et al. (2020)) Death rate in the US is set to 1.9% according to Johns-Hopkins Mortality Analysis.]

This idea is based on an assumption that the death number is counted more accurately and the test resources are firstly distributed to more severely ill patients. Hopefully it can help us reduce the effect of test capacity. However, even with the average death rate and the transformation above, the real situation is too complicated to be explained in this simple setup. We have not considered many other factors such as average hospitalization time, hospital capacity and age distribution of each state etc. Therefore, although the idea is quite brilliant, it was too hard to practice, so we finally discarded it.

3.2.3 *Interpretation*

The new model has a better interpretation of how a lockdown works. Lockdown is no longer a single binary variable that is either on or off but has been broken into several measurable mobility of specific areas. That is, a measurable level of lockdown can be obtained and then we can have a quantitative description of how the spread of COVID-19 is associated with lockdown levels. Compared to the old model here, the new model is more specific in that it does not depend on the continuous time variable which does not show strong relation with the growth of cumulative positive cases. Moreover, based on the mobility data, we know how the lockdowns were actually implemented, therefore our regression now depends on the real situation of each state instead of the announcement and statement made by local governments.

3.2.4 *Result*

Based on the second model, we have the outcome of the United States which had officially declared statewide lockdowns around March 2020, and we choose the data of 8 months from March 20, 2020 to Nov 20, 2020, after filtering the States which satisfying the homoscedasticity and other assumptions(which is 31 out of 51 states), the specific coefficients can be checked in Table.2.

From the result in Figure.1, we can find some commonalities among the same coefficient of different states. Based on our second model, the common negative coefficient indicates that increasing mobility of a certain place has a positive effect of slowing down daily new cases. Common positive coefficients vice versa. From the figure mentioned before, we can clearly tell that there are strong common positive coefficients for workplace and retail-recreation. For grocery-pharmacy, the coefficients for different states are closely all negative. While the coefficient for parks is negative and relatively much smaller compare to other coefficients. Overall result shows that increasing mobility in places like grocery-pharmacy and parks slows down the spread of COVID-19. Gathering around workplace, retail and recreation places will accelerate the spread. The result is reasonable according to our background knowledge. Staying in workplace increases physical connections therefore speeds up the spread. While parks, or more generally, outdoors with good ventilation are not conducive to the spread of virus. Therefore, we can conclude that for lockdown policies, it is crucial to reduce workplace and recreation mobility and keep people inside.

Besides, we have the outcome on regions of Italy which had declared a nationwide lockdown on Mar 3, 2020, and we choose the same length of time, i.e. 8 months from March 20, 2020 to Nov 20, 2020. After our experiment, the specific coefficients can be checked in Table.3.

From Figure.2, the result here shows some differences from that of the United States. Lockdowns did not show a significant or even positive impact on slowing down the spread of COVID-19. For instance, negative coefficients along with large magnitude on the grocery-pharmacy variable were expected according to the model on the United States, but no clear evidence shows reducing the mobility of grocery-pharmacy slows down the increase of the positive cases in Italy. Several factors might have resulted in this. Our conjecture is related to lower testing capability and relatively steady growth of positive cases in Italy. Not like the United States, after the first outbreak in March, the spread of virus had been stabilized until November while the growth of cases is still surging in the United States. Therefore, we can roughly conclude that large-scope lockdown has a powerful effect only on the regions which are experiencing an outbreak with surging cases.

Note that after Box-Cox transformation, because of the different λ we have picked for different states and regions, the coefficients of the explanatory variable from different states are not comparable. However, notice that for all λ, $Y_i^{(\lambda)}$ is monotonically increasing for all $Y_i > 0$. This implies that a positive coefficient tells us an increase per unit for a certain explanatory variable will lead to an increase on $Y_i^{(\lambda)}$, thus, an increase on Y_i. Also, the comparison between any two coefficients of mobility of two different areas in each single state or region is worth studying, because we only did the transformation to the response variable which means the explanatory variables are still in scale for each single state or region.

3.3 Model 3: Multiple Linear Regression Model for Google Trends

3.3.1 *Model setup*

Later, with this relatively integrated model, we can apply this model to other data that can help measure lockdown levels. A generalized lockdown can be not only physical but also psychological, for example, people's attitude towards some certain CDC suggestions can affect the spread of Covid-19. Here we choose trends, i.e. google search index with key words: face mask, hand sanitizer, quarantine, and COVID-19 testing to help measure this psychological lockdown. These terms, to a certain extent, can reflect the degree of the public's awareness of lockdown and other measures. This model is still a Multiple Linear Regression Model like Model 2, where

- $\boldsymbol{Y}$ denotes new positive cases matrix by days in the specified state or region.

- $\boldsymbol{X}$ denotes, instead, explanatory variable matrix with elements; $X_{i,1}, X_{i,2}, \ldots, X_{i,4}$ corresponding to the trends of term "face mask", "hand sanitizer", "quarantine" and "COVID-19 testing", respectively.

And the model modifications are the same as mentioned before, here we also used Box-Cox transformation in this multiple linear regression model. We are here just to verify the universality of this model, this section we simply choose 4 states, which are Michigan, California, New York, Illinois and Connecticut mentioned before.

3.3.2 *Result*

Successfully, we got the results in the chosen states, which can be seen in Table.4 and Figure.3. The results tell us this comprehensive model can be modified to certain forms to do research on many aspects of COVID-19 including measuring the lockdown level, so it is universal to analysis the pandemic and make more related policies by government, for example, raise the awareness of the public towards CDC suggestions such as hand washing and wearing a mask to slow down the spread of COVID-19.

4 Discussion

4.1 Model 1

In the first model above we assume that whether the states are locked down or not is the only factor besides time that affects the rate of increase of confirmed case of COVID-19, but in reality, there are so many other factors that can also influence the spread of the COVID-19, such as transportation, hospital capability, and awareness of wearing masks, etc. Neglecting these factors may result in predictable changes in the regression.

Moreover, in this model, we assumed that time is one of the factors that influence the cumulative cases and we used time as a continuous variable in the model before, but we did not get a satisfactory result. Therefore, in our new model, we removed time variable from the model and use the location mobility as new explanatory variables. Luckily, we got a well-fitted model.

4.2 Model 2

The second model is much more specific and has fitted the reality much better than the first one. In the explanatory variables part, we break lockdown into mobility numbers in different places, measuring lockdown level by mobility change. We solved the multicollinearity problem by variable choosing. In the response variable part, we considered a time lag issue and we fixed it by adding a reasonable time lag for 7-14 days. We noticed a latent risk of undercounting daily new cases number, so we tried to compute a weighted sum with a theoretical positive increase which is obtained from the new death number. We also use a Box-cox transformation to fix the normality assumption and equal-variance assumption without much effect on the interpretability of explanatory variables.

4.3 Model 3

The third model is based on the same idea of breaking down lockdown factors. Lockdown contains many aspects, not only a physical lockdown which is measurable through mobility data, but also a psychological lockdown which may refer to people's awareness of wearing a mask, social distancing, etc. We think that Google trends data is a good way to estimate that abstract lockdown level. But because trends data varies much more than mobility data in all ways, e.g, trends for face mask increases by 30 times from January to April, and they bring more multicollinearity problem to the system if we mixed mobility and trends together. We decided to build a new model for trends data only.

5 Future

We believe our data is though not perfect, however, is well fitted. Model 2, which we believe is optimal, described the behavior of lockdowns in different parameters with acceptable accuracy. To answer our question, we may look at or compare the coefficients of mobility of each area in each single state or region. The model is not built to make predictions, even though it may work with certain accuracy. To have future work on the model, we may increase the prediction accuracy of the model by reducing the variance of the error. Moreover, the death rate inference idea was brilliant, but it brought us many unsolved or unsolvable problems based on our current knowledge. However, it will be a good aspect to continue working on the project in the future.

6 List of items

Item 1: Page 1-6
Item 2: Page 7, we do not have a copy of rubric because we do not know how to copy the entire table to our Latex document.
Item 3: Page 8
Item 4: Page 1-6
Item 5: Page 1-6
Item 6: Page 1, Lines 1-13
Item 7: Page 1, Lines 1-13
Item 8: Page 1, Lines 1-13
Item 9: Page 1, Lines 1-13
Item 10: Page 1, Lines 1-13
Item 11: Page 1-2, Lines 14-50
Item 12: Page 1-2, Lines 14-50; Page 6, Lines 207-231
Item 13: Page 3-4, Lines 108-136
Item 14: Page 2-6. Lines 51-205
Item 15: Page 2-5
Item 16: Page 3 line 110: also full assumption diagnosis are stated in code(page 1 line 5)
Item 17: Model 1: Page 2 line 65; Model 2: Page 3 line 140; Model 3: Page 6 line 225
Item 18: Page 10-13
Item 19: Model 1: Page 2 line 65-75; Model 2: Page 3 line 140-150; Model 3:Page 6 line 200
Item 20: Page 3, Lines 89-108
Item 21: Page 3 to 4, Lines 109-136
Item 22: Page 4-5, Lines 145-192; Page 6 200-205
Item 23: Page 5, Lines 151-161, 162-174
Item 24: Page 5 line 175 line 155 line 170
Item 25: Page 5 line 155 line 170

7 Reflection

7.1 What you have learned (or not learned) about data analysis through this project?

We have learned a lot from the course. First, we learned how to raise a question from the perspective of a statistician. We need to take serious considerations of every aspect of the question that might affect. Moreover, every single small thing we have neglected might cause serious unstableness or inaccuracy of the answer to the model, then finally have an unconvincing answer to the raised question. Second, there is almost no data that are ideal and perfect to build a model, we need to try our best to modify the data to make it workable for our question without losing any essential information. Additionally, after modifying the data, the relationship between original data and modified data must be clear in order to have enough reference to change the model.

7.2 What would you recommend to change/improve about the project for future students?

The meetings with professor were quite helpful, but they were too short. We had many questions that were not asked in the meeting. Moreover, professor's suggestions mainly focused on the reasonability or the logistic of building the model, but sometimes more detailed suggestions on each step of the project would help more.

7.3 Do you think the skills that you have learned in this project will be useful to you in the future?

The skills we have learned are quite helpful. Even though our project might be very primitive and simple, we know how to do a more complicated one based on what we have learned and done in this final project. We have learned a lot of practical and useful tricks or techniques to process the non-ideal data. After doing this final project, we will be able to build much more complicated model in other courses or when we are working in the future.

References

Box, George EP and David R Cox, “An analysis of transformations,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 1964, *26* (2), 211–243.

Lauer, Stephen A, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler, “The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application,” *Annals of internal medicine*, 2020, *172* (9), 577–582.

Li, Pengfei, “Box-Cox transformations: an overview,” *presentation, http://www. stat. uconn. edu/˜ studentjournal/index_files/pengfi_s05. pdf*, 2005.

Leonhard H, Niel H, Philip N, et al. Handbook of infectious disease data analysis [M]. London: CRC Press, 2019.

A Appendix

A.1 Result of Model 1

State	MI	CA	NY	IL	CT
β_1	847.6928	1993.1970	2526.2267	1403.9872	654.6778
β_2	74.47816	5420.14367	65105.26447	-3408.07660	-3240.41149

Table 1: The regression coefficients by initial model in certain states in United States

A.2 Result of Model 2

state	retail and recreation	grocery and pharmacy	parks	workplaces
Alabama	0.047015	-0.04019	0.001076	0.016043
Alaska	0.190712	-0.20959	-0.01537	0.103702
Arkansas	0.406515	-0.42423	0.007843	0.170176
Colorado	0.003822	-0.00163	-0.001	0.00015
Delaware	0.015078	-0.01637	-0.00377	0.005004
Hawaii	0.503532	-0.39598	-0.14817	0.146483
Idaho	0.210254	-0.17781	-0.00533	0.073385
Illinois	0.004746	-0.00363	-0.00152	0.003537
Indiana	0.000849	-0.0014	-7.7E-05	0.000634
Kansas	0.024898	-0.02833	-0.00179	0.010425
Kentucky	0.018962	-0.02219	-0.0004	0.006769
Maine	0.000788	0.001164	-0.00057	0.001407
Maryland	0.000566	-0.00087	-0.00017	0.000364
Massachusetts	-0.00939	-0.00526	-0.00726	0.009108
Michigan	0.023779	-0.03618	-0.00396	0.019484
Mississippi	0.020343	-0.02447	0.000995	0.005088
Missouri	0.047171	-0.06057	-0.00255	0.025677
Montana	0.20948	-0.36266	0.013639	0.205447
Nevada	0.00641	-0.00342	-0.0009	0.001339
New Hampshire	0.002915	-0.00115	-0.00123	0.001502
New Jersey	-0.00076	-0.00095	-0.00202	0.003668
New York	-9.5E-05	-0.00012	-0.00011	0.000184
Ohio	0.002014	-0.00208	-0.00015	0.001046
Oklahoma	0.24269	-0.30604	-0.0129	0.104018
Pennsylvania	0.000515	-0.00088	-0.00015	0.000472
Tennessee	0.011557	-0.01531	0.000541	0.004454
Utah	0.010112	-0.00428	-0.00052	0.003399
Virginia	0.109222	-0.07236	-0.01119	0.023194
West Virginia	0.042797	-0.06688	-0.00102	0.033426
Wisconsin	0.014561	-0.00721	-0.00133	0.005064
Wyoming	0.02679	-0.04774	0.001685	0.028813

Table 2: The regression coefficients by Model 2 in all of the states in United States

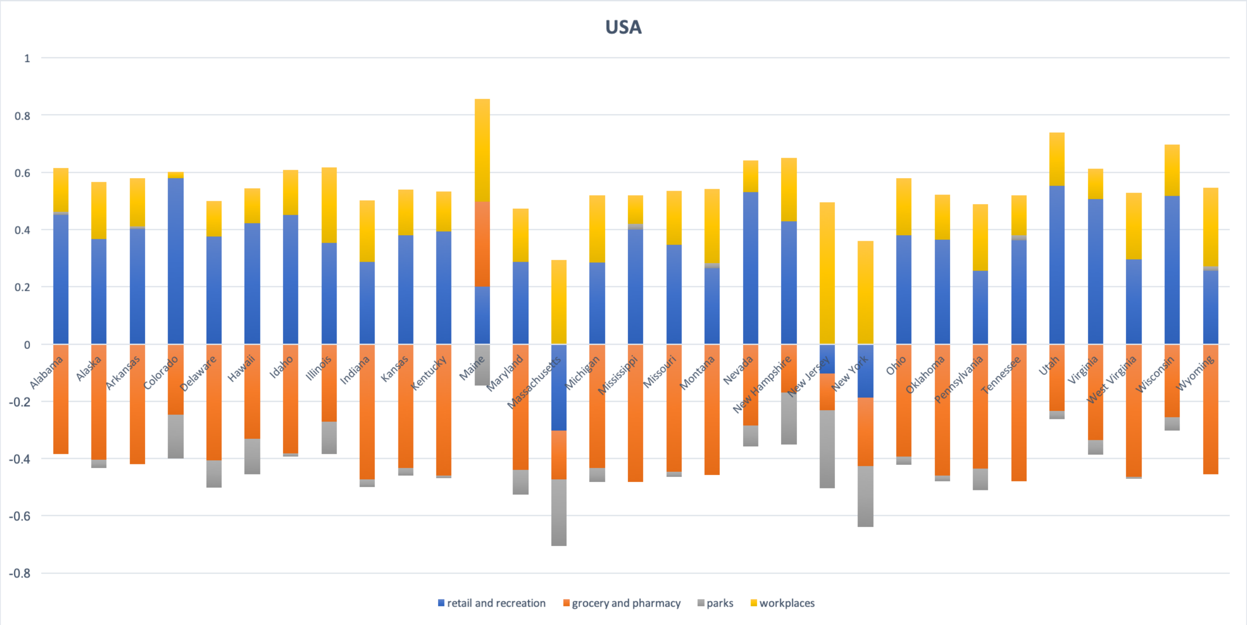

Figure 1: Percentage accumulative histogram of USA coefficients from March 20, 2020 to Nov 20, 2020

region	retail and recreation	grocery and pharmacy	parks	workplaces
Aosta	-0.0581	0.044067	-0.00451	0.036514
Apulia	-0.0687	0.087238	0.000918	0.012381
Basilicata	-0.08409	0.068392	0.003728	0.073461
Calabria	-0.02776	0.038472	-0.00154	0.010022
Campania	-0.00753	0.054509	-0.0088	-0.0188
Emilia-Romagna	0.009154	0.004196	-0.00742	-0.00274
Liguria	-0.04139	0.077614	-0.01195	-0.00074
Marche	-0.02327	0.05823	-0.01222	0.004976
Sardinia	0.010123	0.045686	-0.01017	0.006673
Sicily	-0.03436	0.05256	-0.00304	0.02375
Trentino-South Tyrol	-0.09065	0.041598	0.011234	0.055315
Tuscany	0.006223	0.031612	-0.01049	-0.00935
Veneto	0.056071	0.005932	-0.02701	-0.01799

Table 3: The regression coefficients by Model 2 in Italy

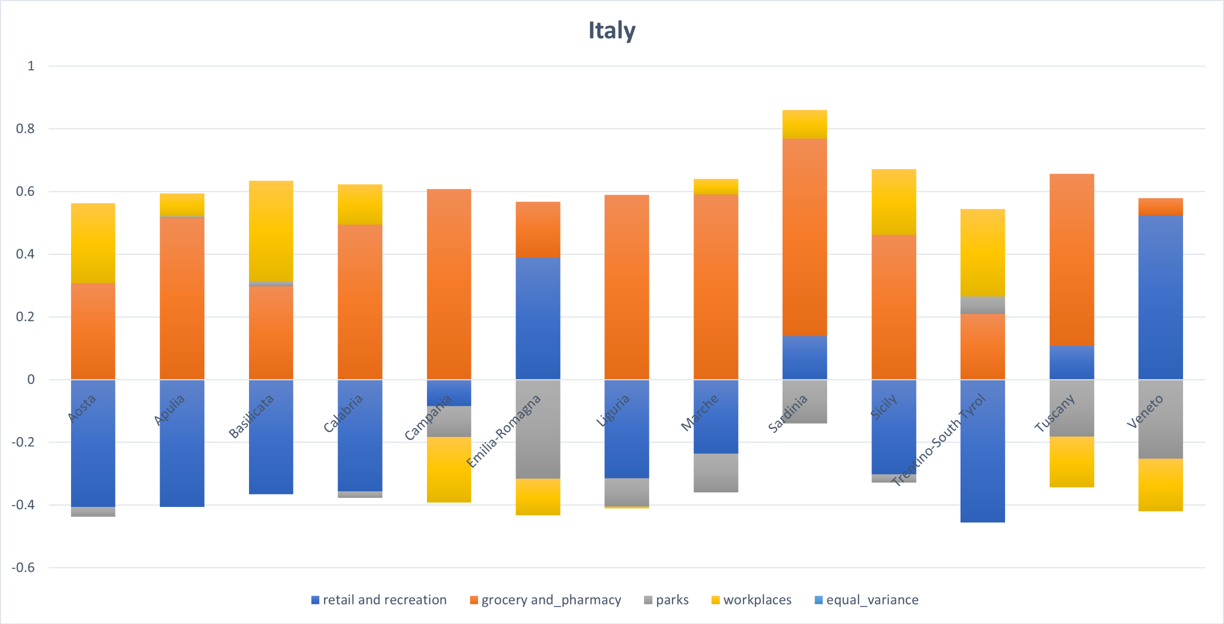

Figure 2: Percentage accumulative histogram of Italy coefficients from March 20, 2020 to Nov 20, 2020

A.3 Result of Model 3

state	face mask	hand sanitizer	quarantine	testing
MI	-0.06234	-0.31431	0.363457	0.269214
CA	0.004299	-0.02825	-0.06687	0.082406
IL	0.016031	-0.03333	0.012848	0.021068
CT	0.187223	0.077524	-0.03533	0.129559

Table 4: The regression coefficients by Model 3 and trends data in certain states in United States

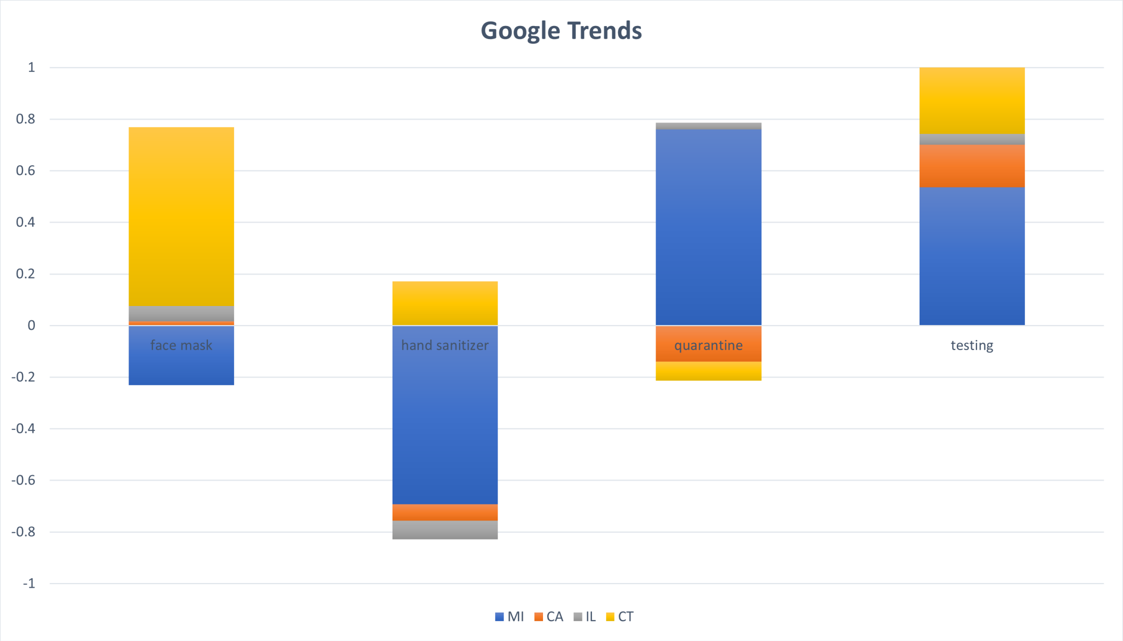

Figure 3: Percentage accumulative histogram of USA trends coefficients from March 20, 2020 to Nov 20, 2020

B Other experiment results

state	retail and recreation	grocery and pharmacy	parks	workplaces
Alabama	2.839805	3.164085	2.202272	1.486083
Alaska	2.397635	2.710797	1.548608	1.852262
Arizona	4.959155	3.717312	2.224	1.949202
Arkansas	2.960875	2.311592	1.562371	1.458114
California	2.386698	1.671014	1.919988	1.494819
Colorado	7.945614	5.603985	2.674012	1.44923
Connecticut	4.338273	3.88283	1.603557	1.41529
Delaware	9.875277	10.96974	1.31307	1.725908
District of Columbia	8.212048	5.175175	4.076157	5.335742
Florida	9.134762	11.47824	4.337405	2.007572
Georgia	4.632189	4.259492	1.79242	1.290763
Hawaii	5.285267	7.596995	5.077938	3.770233
Idaho	3.710907	3.273959	1.581553	1.278048
Illinois	3.96386	3.444039	1.187159	1.67438
Indiana	3.989192	4.320005	1.04202	1.703784

Table 5: VIFs (variance inflation factor) in some states

C Codes

All codes and data are available on the Github.
USA cases:
`https://github.com/ShermanSteinke/STAT601/tree/master`
Italy cases:
`https://github.com/Mengqi411/stat601_final_project/blob/main/Italy%20cases_without0.R`