ISYE 6420 – Project

ekhan33

23April 2023

1. Introduction

Covid19 presented an unprecedented challenge for humans around the world. Policymakers were challenged by the ever growing, sometimes contradicting research on how its spread can be contained. Most popular strategies included government restrictions like “Stay at home”, “Workplace closing”, “Public event cancellations”. In this analysis these popular strategies are analyzed to estimate their impact on containing covid19 spread. It is very valuable for policymakers and even individuals to understand which methods are effective in containing this disease. Bayesian regression is explored in this analysis to understand the impact of policies on the forward looking average of newly confirmed covid-19 patients. Another hypothesis is that the effectiveness of different strategies may also have varied effects in different time periods. Therefore, two separate time periods are analyzed for this assessment.

Wibbens et al [1] have previously worked on a Bayesian based impact analysis for different government restrictions imposed during the period of interest. Their work focuses on the entire dataset and provides a cross-geographical perspective. This project focuses on the countries that were most affected by COVID19 (in terms of volume) and leverages multiple linear regression (MLR) technique to gauge and quantify this impact. Bayesian results are then compared with a Frequentist MLR model. There is another aspect closely linked with effectiveness, the primary variant associated with a covid “wave”. The differences in policy effectiveness are thus explored in different time periods or “covid waves”.

2. Data Collation and Description

Data used for this analysis is open source and available on Google Raw Database. (Ref: https://health.google.com/covid-19/open-data/raw-data).

Oxford COVID-19 government response tracker is the original source for data related to government restrictions. Different data sources are used to fetch epidemiology dataset, for different regions:

US: COVID Tracking Project
IN: Wikipedia
FR: Robert Koch Institute
BR: Brazil Ministério da Saúde
DE: Robert Koch Institute

Complete information on the region level sources available on the link referenced above.

2.1 Data Description

The epidemiology dataset (1,25,25,825 rows) contains day wise information about infections at a region level. It contains both data for the current date and cumulative metrics. This analysis will be focused on analyzing the impact of policies on newly confirmed cases. Therefore "new_confirmed" will be the primary metric from this dataset. The "key" is the unique identifying key of the region, the information is available at both country level (US) and more granular levels (US_CA).

Table 1: Epidemiology data description

Epidemiology Dataset

Name	Type	Description
date	string	ISO 8601 date (YYYY-MM-DD) of the datapoint
key	string	Unique string identifying the region
new_confirmed	integer	Count of new cases confirmed after positive test on this date
new_deceased	integer	Count of new deaths from a positive COVID-19 case on this date
new_recovered	integer	Count of new recoveries from a positive COVID-19 case on this date
new_tested	integer	Count of new COVID-19 tests performed on this date
cumulative_confirmed	integer	Cumulative sum of cases confirmed after positive test to date
cumulative_deceased	integer	Cumulative sum of deaths from a positive COVID-19 case to date
cumulative_recovered	integer	Cumulative sum of recoveries from a positive COVID-19 case to date
cumulative_tested	integer	Cumulative sum of COVID-19 tests performed to date

The government response dataset (3,03,969 rows) contains information on the restrictions or policies used by government in different regions. There is an indicator (0-3) for each policy that shows the extent to which the policy was advised. Other metrics such as emergency investment made in healthcare, vaccines, fiscal measures an international support is also available (in USD).

Table 2: Government Response data description

Government Response Dataset

Name	Type	Description
date	`string`	ISO 8601 date (YYYY-MM-DD) of the datapoint
key	`string`	Unique string identifying the region
school_closing	`integer [0–3]`	Schools are closed
workplace_closing	`integer [0–3]`	Workplaces are closed
cancel_public_events	`integer [0–3]`	Public events have been cancelled
restrictions_on_gatherings	`integer [0–3]`	Gatherings of non-household members are restricted
public_transport_closing	`integer [0–3]`	Public transport is not operational
stay_at_home_requirements	`integer [0–3]`	Self-quarantine at home is mandated for everyone
restrictions_on_internal_movement	`integer [0–3]`	Travel within country is restricted
international_travel_controls	`integer [0–3]`	International travel is restricted
income_support	`integer [USD]`	Value of fiscal stimuli, including spending or tax cuts
debt_relief	`integer [0–3]`	Debt/contract relief for households
fiscal_measures	`integer [USD]`	Value of fiscal stimuli, including spending or tax cuts
international_support	`integer [USD]`	Giving international support to other countries
public_information_campaigns	`integer [0–2]`	Government has launched public information campaigns
testing_policy	`integer [0–3]`	Country-wide COVID-19 testing policy
contact_tracing	`integer [0–2]`	Country-wide contact tracing policy
emergency_investment_in_healthcare	`integer [USD]`	Emergency funding allocated to healthcare
investment_in_vaccines	`integer [USD]`	Emergency funding allocated to vaccine research
facial_coverings	`integer [0–4]`	Policies on the use of facial coverings outside the home
vaccination_policy	`integer [0–5]`	Policies for vaccine delivery for different groups
stringency_index	`double [0–100]`	Overall stringency index

2.2 Data Preparation

The datasets described above were linked based on date and location_key. Data was filtered for the regions where government_restriction information was available. There was significant variation in the spread of COVID19 in different countries based on the Epidemiology dataset However, certain countries overall were affected more in terms of volume. The five regions, US, IN, FR, BR and DE account for 30% of the covid cases worldwide (Fig.1). For this analysis, we are limiting our dataset to include the 5 most affected countries as there are higher chances to see observable changes with changes in the policy. To control the bias introduced by significant fluctuations in rate of change due to other external factors (like the virus variant) that have a higher impact, the time period (Fig. 2.) where the new confirmed cases were relatively stable is considered.

Figure 2: Daywise new confirmed cases

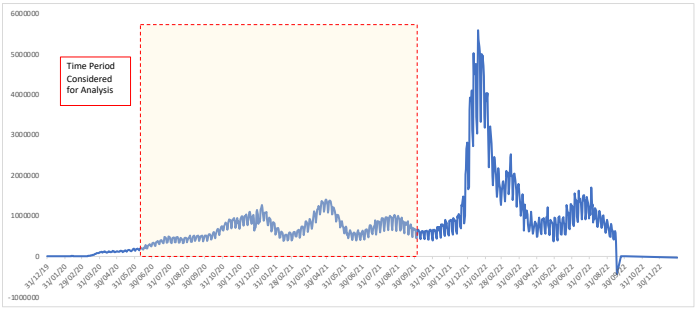


Figure 1: Location wise total covid cases

location_key	new_confirmed	new_deceased	share	cum_share
US	91790598.0	988028.0	11.260519	11.260519
IN	44516479.0	528250.0	5.461111	16.721630
FR	35203157.0	131288.0	4.318588	21.040219
BR	34581186.0	685203.0	4.242287	25.282506
DE	32604993.0	148728.0	3.999856	29.282362

It is important to understand that any policy change made or sustained will likely result in a delayed observable impact. Therefore, the target variable is defined as the forward-looking moving average of confirmed cases of next 7 days. *Scaling*: The dataset contains columns at very different scales and ranges. To ensure that the data in on the same scale, min-max scaler is used from sklearn library. The impact of the outliers therefore can potentially be reduced. This also enables easier interpretation of the intercept and coefficient values determined by the model. *Variable Selection*: As the model is to be developed with multiple input variables it is important to ensure that these variables are not correlated. That is, multicollinearity is avoided. VIF is therefore estimated for each variable. As a rule of thumb, >5 VIF is considered as high. The variables "public_information_campaigns" and "stringency_index" are thus removed from the predictor dataset (Fig. 3.).

	feature	VIF
0	school_closing	3.137943
1	workplace_closing	4.829982
2	cancel_public_events	4.344164
3	restrictions_on_gatherings	4.777965
4	public_transport_closing	4.219948
5	stay_at_home_requirements	3.707449
6	restrictions_on_internal_movement	3.025291
7	international_travel_controls	2.982962
8	income_support	2.759152
9	debt_relief	2.382865
10	fiscal_measures	1.088990
11	international_support	1.931897
12	public_information_campaigns	95.838921
13	testing_policy	1.855051
14	contact_tracing	1.985205
15	emergency_investment_in_healthcare	1.010287
16	investment_in_vaccines	2.028485
17	facial_coverings	2.520270
18	vaccination_policy	2.898647
19	stringency_index	20.399198

Figure 3: VIF Score of predictor variables

3. Analysis and Results

3.1 Methodology

Frequentist Method

There may be an impact on forward looking 7 day moving average of COVID19 infections by any or combination multiple policies being considered. Therefore, Multiple linear regression is primarily used. Regression analysis allows us to analyze and quantify the linear relationship between the independent and the dependent variables where the random error (ε) is assumed to be normal and have constant variance.

Y = β0 + β 1*xi.. + ε

The cost function is used as a basis to optimize the regression coefficient and minimize error. Linear regression makes use of the "Mean Squared Error" or MSE as the cost function. MSE is the average of squared error between the predicted value and actual value.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

Bayesian Regression

In Bayesian regression, Bayesian principles are used to estimate the model parameters. Priors are specified for model parameters and based on posteriors, the best fit is defined. Data is in the form $(x_{i1}, x_{i2}, \dots x_{ik}, y_i), i = 1 \dots n$
The model is defined as:

$$Y_i = \beta_0 + \beta_1.x_{i1} + \beta_2.x_{i2} + \cdots + \beta_k.x_{ik} + \epsilon_i$$
$$Where\ \epsilon_i\ is\ iid \sim N(0, \sigma); k + 1 = p$$

Bayesian approach allows us to model uncertainty in parameters. The uncertainty in the predictions can also be analyzed.

3.2 Frequentist Vs Bayesian

On analyzing the impact of policies on the forward looking moving average of new covid patients, the variables linked to "cancel_public_events", "income_support", "school_closing", "workplace_closing", "debt_relief", "international_support", "fiscal_measures" and "emergency_investment in healthcare" have a non-zero coefficients which negatively effects the number of new patients on an average. While the other variables may have a long-term impact on rising cases, there is no observable immediate improvement. The frequentist and Bayesian MLR coefficients estimated for most variables are similar (Table 3). This behavior is as expected since non-informative priors were used for this analysis.

Table 3: Frequentist and Bayesian MLR Coefficients

Variable name	Coeffiicients	Bayesian Coefficients	Frequentist Coefficients
cancel_public_events	beta1	-0.137	-0.13728
income_support	beta2	-0.02	-0.02039
restrictions_on_gatherings	beta3	0.037	0.03664
international_travel_controls	beta4	0.004	0.00374
testing_policy	beta5	0.002	0.00223
school_closing	beta6	-0.023	-0.02286
contact_tracing	beta7	0.028	0.02809
workplace_closing	beta8	-0.022	-0.02169
debt_relief	beta10	-0.066	-0.06598
international_support	beta11	-0.023	-0.02498
investment_in_vaccines	beta12	0.01	0.01123
fiscal_measures	beta13	-0.008	-0.00838
emergency_investment_in_healthcare	beta14	-0.013	-0.01324
public_transport_closing	beta15	0.018	0.01760
restrictions_on_internal_movement	beta16	0.041	0.04125
facial_coverings	beta17	0.019	0.01944
vaccination_policy	beta18	0.008	0.00808
stay_at_home_requirements	beta19	0.071	0.07115

Figure 4: Distribution of Bayesian Coefficients

On closely analysing distribution of Bayesian coefficients it is observed that beta coefficients (Fig. 6.) for variables “cancel_public_events”, “debt_relief”, “income_suport” and “testing_policy” do not contain “0” in their 95% credible set and vary between negative values. This is indicative of that an observable impact exists with the dependant variable.

3.3 Variation in impact based on time-period considered

To understand if there is any difference in the impact of polices within different timeframes, 2021 and 2022 data was analyzed again separately. On running Bayesian MLR, it is observed that there are substantial differences in the observed impact of policies within the two timeframes (Fig. 5 & Fig. 6). While there is higher impact of variables like public_transport_closing, fiscal_measures, international_support, workplace_closing, testing_policy, international_travel_controls, cancel_public events in 2020, impact of variables stay_at_home_requirements, debt_relief, school_closing is more significant in 2021. In general, government restrictions seem more effective in 2020 compared to 2021.

Figure 5: Bayesian Coefficients on 2020 Data

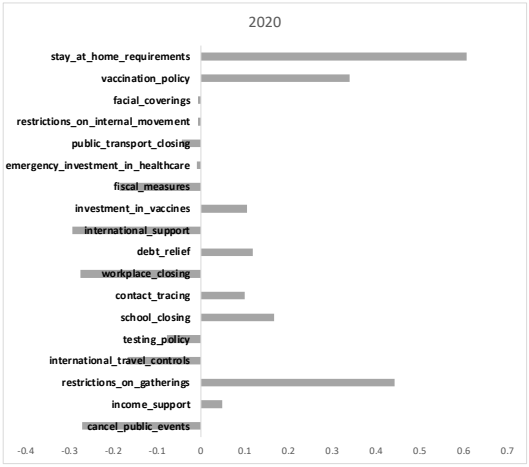
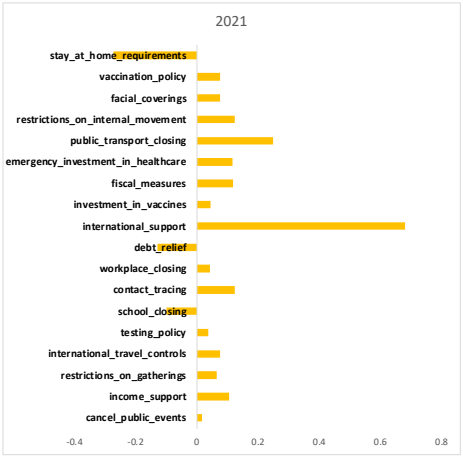


Figure 6: Bayesian Coefficients on 2021 data



4. Conclusion

Based on the results obtained in the previous section, it can be observed that certain government restrictions during the covid19 outbreak impacted the spread visibly. The coefficient credible sets for variables associated with policies like “cancel_public_events”, “debt_relief”, “income_suport” and “testing_policy” indicate that these factors have an observable impact within a short duration on reducing the spread. Both Bayesian and Frequentist models yielded similar coefficients for the variables under consideration. There were however, significant differences in regression coefficients when different time periods were considered. In 2020, policies on public_transport_closing, fiscal_measures, international_support, workplace_closing, testing_policy, international_travel_controls, cancel_public have a more significant coefficient. In 2021, stay_at_home_requirements, debt_relief, school_closing is more significant in 2021. In general, government restrictions seem more effective in 2020 compared to 2021.

REFERENCE

[1] Wibbens PD, Koo WW-Y, McGahan AM (2020) Which COVID policies are most effective? A Bayesian analysis of COVID-19 by jurisdiction. PLoS ONE 15(12): e0244177. https://doi.org/10.1371/journal.pone.0244177