HS4300 Applied Economics

Jul-Dec 2020

Term Project

Effect of Different Factors on COVID-19 in Different States of the USA

CH17B009 - Nandini Vasuki Garimella
CS17B045 - Vamsi Krishna Valluru
ME17B114 - S Tarun Prasad
ME17B158 - Omkar Nath
ME17B170 - Uma T V

Acknowledgement

We would like to express gratitude to our course instructor, Dr. Santosh Kumar Sahoo, for giving us the opportunity to apply what we learnt in our classes and research paper discussions to a real-life case study through this project. This project was a great experience for all of us. We learnt many essential skills like finding the right data in web sources, reading various sources to make predictions of the trends, visualizing and analyzing data through Regression techniques, critical thinking to explain the observed values of coefficients for various factors, and lots more.

We would also like to thank all the authors, researchers and economists for the work they have done, which we referred to during our project.

We are grateful to the institute for their constant support and adaptability to keep our education happening smoothly during these unprecedented times.

Finally, we would like to thank our families for all their love and support.

Table of Contents

Abstract

Life as we know in 2020 has been adversely affected by the virus COVID-19 which has led to a global pandemic. The virus has already claimed over a million lives, and is claiming more every day. It is with this backdrop that we have chosen our project, to identify the effect various factors have on the number of COVID-19 deaths. Specifically, 7 factors have been chosen, which are Age, number of Hospital Beds, Political Affiliation, Education, Income, Population and Temperature. The data considered is for different states of the USA. A linear regression model is used to fit the data, and identify the dependencies of the number of COVID-19 deaths on the various chosen factors. This is followed by an analysis of the observed results, where we try to explain the reasons for the results produced. The hope is that a better understanding of these factors will better help us to combat this virus.

Introduction

In the year 2020, the world was hit with a pandemic, caused due to a new strain of the coronavirus, formally known as COVID-19. This pushed the world into a crisis that it was ill-equipped to deal with. The effects of the pandemic are the worst the world has ever seen since over 70 years. It is unprecedented in its global reach and impact, posing formidable challenges to policymakers and to the empirical analysis of its direct and indirect effects within the interconnected global economy. The global recession will be long lasting, with no country escaping its impact regardless of their mitigation strategy.

Two fundamental strategies have been taken globally against the virus, one focused on mitigating but not necessarily stopping the virus spread and the other relying on more stringent measures to suppress and reverse the growth trajectories. While the USA initially implemented the former strategy, it has since shifted towards the more stringent suppression strategy.

The aspect that makes the virus so scary is its high spread rate, which has led to a high rise in the number of cases over the globe, in spite of various precautionary measures being taken. As such it becomes important to understand the trends of the number of cases, as it helps to our understanding of the containment of COVID-19, and also provides insights into future prevention work.

In this paper, we hope to do just that, by analyzing the effects of various factors such as income, education, etc. on the coronavirus. The factors chosen here cover both social and economic aspects of people. The economic factors contributing to the accessibility to better healthcare or safety measures the people may have available. The social factors such as education also become equally important as over the last few months, efforts have been made to have the community take precautionary measures against the virus, with the hope to combat its overall spread.

The effects on the virus are considered using the number of coronavirus cases as a benchmark for comparison. The basis for comparison of this is the various states of the USA for 2 reasons. Firstly, they provide an accurate segregation of various groups of interest, representing differing levels amongst the factors of consideration. Secondly, there is abundant data easily available which is accurate and consistent across states for us to use in our analysis.

By doing this analysis, we hope to answer two key questions in our paper. Firstly, whether there exists a correlation between the various chosen factors and the number of coronavirus cases or if there is no effect at all, and secondly, as to what the factor of relation is. We answer these by using various data analytics tools on the cross-sectional data available to us.

In this manner, we hope to be able to get an understanding of the effects of certain variables on the coronavirus spread, which can hopefully be used to better understand and contain the virus.

Factors of Dependence

For our analysis, we choose 7 factors which we think are important to understanding the effect the COVID-19 virus has on various demographics. Understanding the effect of these factors can help us to better develop ways to combat the virus. The 7 factors we have chosen are as follows:

1. Age
   The median age of the population of the state
2. Beds
   The number of beds per 1000 people in the state
3. Democrats
   The percentage of democrats in the state
4. Education
   The percentage of the population of the state that received the Bachelors
5. Income
   The average income of the population of the state in dollars
6. Population
   The total population of the state
7. Temperature
   The average temperature of the state in degree Celsius

For the analysis, these factors are used as dependent variables, to try and predict the number of COVID-19 deaths. Here we consider the number of deaths and not the number of COVID-19 cases as we are more concerned with the number of deaths caused by the virus, as that is the end result which everyone is trying to avoid. An example to understand why this is so is the common cold. While the common cold is very contagious, it is not lethal which is why we are not very worried about it.

Analysis Methodology

To approach this problem, we need to quantify the relationships between the various possible causal factors considered and the observed COVID data in a statistical manner. This is required in order to conclude whether the assumptions made on the relationships hold good or if not to identify the true relationships and try to reason why it happens to be so. The methodology we have used here to understand the relationships between the observations and the possible causes is Linear Regression. Linear Regression is a linear approach to modelling the relationship between a scalar output and one or more dependent and independent explanatory variables. It assumes that the output can be computed as a linear combination of these variables and tries to compute what this linear combination is. This is represented mathematically by the equation below where y is a column vector of the various observed values, x a column vector of '1' for bias followed by all the possible causal factors and B a matrix whose elements correspond to the respective proportionality coefficient between each observed value and each element in x.

$$\mathbf{y = B*x}$$

The goal of the process is to find the most optimal values for the elements in the B matrix also known as the weight matrix such that the output y predicted by it using the different data points for x are close as much as possible to the true observed value of y for the corresponding data points. To summarize this, we need to define a function which measures the closeness of the predicted values to the actual values. One of the most commonly used functions for the same is the sum of squared errors. This ensures that the sign of different errors don't cancel out each other while also being differentiable at all values of errors in contrast to absolute error which is not differentiable when the error is zero. The solution to the minimization of this function is given as:

$$\mathbf{B = (x^T x)^{-1} x^T y}$$

Using the computed value for B we can isolate the weights corresponding to different elements in x also known as features and verify the hypotheses of relationships between the output and the different causal factors and also study the order of contribution of one factor over the other and obtain which were the more principal features.

Expected Effects

Before moving on to the quantitative analysis of the data, we would like to do a qualitative analysis of what we expect the dependence of the number of COVID-19 deaths on the various factors to be. This is done before the results of the data analysis, which may or may not match with our postulated correlations.

Age

Since when the COVID-19 spread initially started to happen, it was believed that those who are aged over 60 years, and people who have underlying medical conditions such as diabetes, heart disease, respiratory disease or hypertension are among those who are at greater risk of developing severe or critical illness if infected with the virus. This would automatically translate to a higher number of COVID related deaths.

Expectation: *States with higher median age will see more COVID-19 deaths (positive correlation)*

Beds

The number of beds in a state is an indication of the number of people who can get timely treatment in the event that they are infected with the disease. Now, state populations and sizes may vary drastically across the United States due to which the number of beds in a given state may not give us the insight we seek to gain. Hence, we use the number *of beds per 1000 people* to understand the density of beds in the state. Thus, based on the logic explained above, the expectation is that as the density of beds increases in a state, we will see fewer deaths caused by COVID-19 there.

Expectation: *States with higher number of hospital beds per 1000 people will see fewer COVID-19 deaths (negative correlation)*

Democrats

The popular press has explored the differences among U.S. states in rates of COVID-19 cases, mostly focusing on political party differences, and often mentioning that political party differences in health outcomes are confounded by demographic and socio-economic differences between Democratic areas and Republican areas.

Given that Donald Trump, on numerous occasions, has underplayed the virus and stated that "*The C.D.C. is advising the use of nonmedical cloth face covering as an additional voluntary public health measure. So, it's voluntary. You don't have to do it. They suggested for a period of time, but this is voluntary. I don't think I'm going to be doing it.*" Based on this, we expect to see the virus impacting the republican (red) states more than the democratic (blue) ones, since we expect people in a majority red state to not follow COVID-19 norms and regulations.

Expectation: *States with higher Democratic following will see fewer COVID-19 deaths (negative correlation)*

Education

The expectation is that states with a higher percentage of people with a Bachelor's degree will also see a fewer number of deaths. This is because there exists a general notion where people with an educational background are more aware of the impact a disease may have and hence take the necessary precautions to reduce the spread of disease, especially a highly communicable one. In the case of COVID-19, this involves appropriate social distancing, sanitizing and use of masks to prevent any form of transmission.

Expectation: *States with higher percentage of Bachelor Level Educated people will see fewer COVID-19 deaths (negative correlation)*

Income

Healthcare is traditionally very expensive in the United States and access to it is quite limited. While Medicare (a national health insurance program in the United States, begun in 1966 under the Social Security Administration and now administered by the Centers for Medicare and Medicaid Services) does exist, it is generally available only for people of age 65 or older, younger people with disabilities and people with End Stage Renal Disease (permanent kidney failure requiring dialysis or transplant). Since income directly translates to accessibility of proper healthcare facilities in the event of a COVID-19 infection, the expectation is that the higher income class will see a fewer number of COVID deaths.

Expectation: *States with higher average income will see fewer COVID-19 deaths (negative correlation)*

Population

The virus is believed to spread much faster in more populated areas. This was why social distancing was mandated along with other protective measures like constant sanitizing of hands and wearing of masks. In many parts of the United States, various measures were taken to restrict crowding and lockdowns were enforced in a stage-wise manner across the country to enforce this. This was done in the hope of reducing the speed of the virus spread in the hope of "flattening the curve".

Expectation: *States with large population will see higher number of COVID-19 deaths (positive correlation)*

Temperature

Another belief present early on was that a fewer number of COVID cases and hence deaths would be seen in traditionally hotter areas of the United States. This was because:

(1) It was thought that the virus would not survive at higher temperatures

(2) Immunity and the ability to cope with and fight a foreign body entering the human body increases with temperature

Expectation: *States with higher average temperature will see fewer COVID-19 deaths (negative correlation)*

Data

Following is the raw data used for the analysis, as extracted from the sources mentioned in the References.

State	Age	Beds	Dem %	Education	Income	Population	Temp.	COVID-19 Deaths
Alabama	39.2	3.1	35	24.5	42751	4858979	17.1	3776
Alaska	34.6	2.2	32	29	62185	738432	-3	129
Arizona	37.9	1.9	39	28.4	47454	6828065	15.7	6821
Arkansas	38.3	3.2	38	22	39753	2978204	15.8	2555
California	36.8	1.8	49	32.6	72786	39144818	15.2	19588
Colorado	36.9	1.9	42	39.4	62442	5456574	7.3	3320
Connecticut	41	2	50	38.4	73730	3590886	9.4	5111
Delaware	40.7	2.2	55	31	72857	945934	12.9	779
Florida	42.2	2.6	44	28.5	47429	20271272	21.5	18874
Georgia	36.9	2.4	41	29.9	54091	10214860	17.5	9648
Hawaii	39.2	1.9	51	32	59056	1431603	21.1	246
Idaho	36.6	1.9	32	26.8	42375	1654930	6.9	1014
Illinois	38.3	2.5	48	33.4	63571	12859995	11	13625
Indiana	37.9	2.7	37	25.3	49520	6619680	10.9	6033
Iowa	38.2	3	40	27.7	56303	3123899	8.8	2522
Kansas	36.9	3.3	31	32.3	55479	2911641	12.4	1679
Kentucky	38.9	3.2	43	23.2	43086	4425092	13.1	2014
Louisiana	37.2	3.3	43	23.4	48249	4670724	19.1	6524
Maine	44.9	2.5	47	30.3	45695	1329328	5	220
Maryland	38.8	1.9	55	39	65527	6006401	12.3	4764
Massachusetts	39.4	2.3	56	42.1	78339	6794422	8.8	10874
Michigan	39.8	2.5	47	28.1	47336	9922576	6.9	10035
Minnesota	38.1	2.5	46	34.8	61051	5489594	5.1	3843
Mississippi	37.7	4	42	21.3	35353	2992333	17.4	3879
Missouri	38.7	3.1	42	28.2	48487	6083672	12.5	4353

Montana	39.9	3.3	30	30.7	43911	1032949	5.9	722
Nebraska	36.6	3.6	36	30.6	61299	1896190	9.3	1159
Nevada	38.1	2.1	46	23.7	49820	2890845	9.9	2249
New Hampshire	43	2.1	44	36	57386	1330608	6.6	544
New Jersey	40	2.4	51	38.1	64232	8958013	11.5	17346
New Mexico	38.1	1.8	48	26.9	44531	2085109	11.9	1673
New York	39	2.7	53	35.3	81465	19795791	7.4	34829
North Carolina	38.9	2.1	43	29.9	51524	10042802	15	5410
North Dakota	35.2	4.3	33	28.9	65484	756927	4.7	977
Ohio	39.4	2.8	40	27.2	53442	11614373	10.4	6753
Oklahoma	36.7	2.8	40	24.8	43759	3911338	15.3	1836
Oregon	39.4	1.6	47	32.3	54358	4028977	9.1	973
Pennsylvania	40.8	2.9	46	30.1	56487	12802503	9.3	11031
Rhode Island	40.1	2.1	48	33	53292	1056298	10.1	1400
South Carolina	39.6	2.4	39	27	43125	4896146	16.9	4466
South Dakota	37.1	4.8	37	27.8	56422	858469	7.3	1033
Tennessee	38.8	2.9	36	26.1	48305	6600299	14.2	4781
Texas	34.8	2.3	40	28.7	55976	27469114	18.2	22729
Utah	31	1.8	30	32.5	55724	2995919	9.2	917
Vermont	42.8	2.1	57	36.8	48044	626042	6.1	75
Virginia	38.4	2.1	39	37.6	60271	8382993	12.8	4147
Washington	37.7	1.7	44	34.5	74347	7170351	9.1	2919
West Virginia	42.7	3.8	41	19.9	38244	1844128	11	789
Wisconsin	39.6	2.1	42	29	53663	5771337	6.2	3562
Wyoming	38	3.5	25	26.7	58609	586107	5.6	257

Code

Following is the code written in Python to perform linear regression analysis on the data. The code has been commented for easy understanding.

```
import numpy as np
from sklearn.linear_model import LinearRegression

# Create Input Data:
data = []

file = open('Age.txt', 'r')
for line in file.readlines():
  data.append([float(line[:-1])])
file.close()

file = open('Beds.txt', 'r')
for idx, line in enumerate(file.readlines()):
  data[idx].append(float(line[:-1]))
file.close()

file = open('DemocraticPercentExitPoll.txt', 'r')
for idx, line in enumerate(file.readlines()):
  data[idx].append(float(line[:-1]))
file.close()

file = open('Education.txt', 'r')
for idx, line in enumerate(file.readlines()):
  data[idx].append(float(line[:-1]))
file.close()

file = open('Income.txt', 'r')
for idx, line in enumerate(file.readlines()):
  data[idx].append(float(line[:-1]))
file.close()
```

```
file = open('Population.txt', 'r')
for idx, line in enumerate(file.readlines()):
   data[idx].append(float(line[:-1]))
file.close()

file = open('Temperature.txt', 'r')
for idx, line in enumerate(file.readlines()):
   data[idx].append(float(line[:-1]))
file.close()

data = np.array(data)

# Create Output:
output = []
file = open('CovidData.txt', 'r')
for line in file.readlines():
   output.append(float(line[:-1]))
file.close()

output = np.array(output)

# Linear Regression on input vs output:
model = LinearRegression(normalize=True).fit(data, output)

# Print out the coefficients:
r_sq = model.score(data, output)
print('coefficient of determination: ', r_sq)
print('coefficients of factors: ', model.coef_)
print('intercept: ', model.intercept_)
```

Results

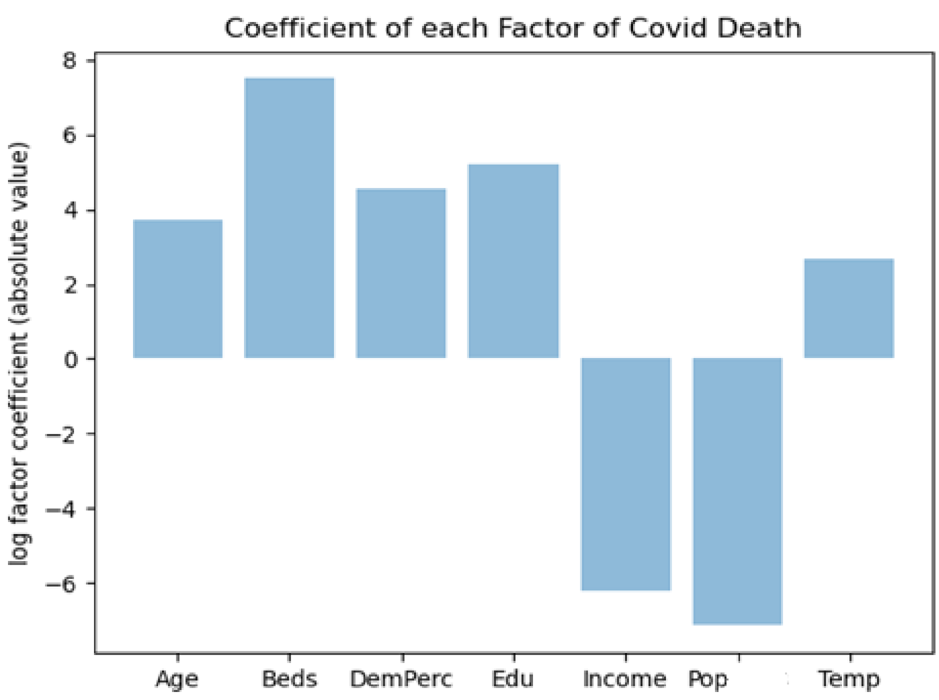
The Coefficients for the various factors as determined via Regression are as follows:

Age	Beds	Democratic Representation	Education	Income	Population	Temperature
4.14	**1797.29**	**9.32**	**183.7**	**0.002**	**0.0008**	**-1.45**

Coefficient of Determination (R-Squared Coefficient) is **0.73**

Intercept is **-15207.3**

Visually, we can see the relative importance of each factor, by plotting the values of coefficients on a log scale as is visible below. A log scale is considered here due to the vast differences between the values of different coefficients.



Discussions

A discussion on the results obtained for the various factors. In this, the results are compared with what was expected, along with a qualitative analysis of why these results might have occurred.

Age

The correlation coefficient for Median Age with the number of COVID-19 deaths is observed as 4.14. This suggests a positive correlation between the two, which is as was hypothesized earlier. The older people tend to be, the weaker their immunity is overall. Additionally, they may be suffering from other medical conditions, which when compounded with COVID-19 could lead to even worse results, often death. Even various governments have suggested older people, especially those above 60 as higher risk groups from COVID-19. One thing to note here is that old people are not more likely to contract the disease as compared to other age groups. Rather we are saying they are more likely to die due to the reasons already mentioned. And since we are taking COVID-19 deaths and not number of cases for our study, the results are thus explained.

One point to consider additionally is that younger age groups are also at a higher risk from COVID-19. However, this can be explained due to three reasons. Firstly, children are much more protected and taken care and thus much less likely to catch the virus in the first place. Secondly, in terms of the overall vulnerability, the older age groups, especially those 60+ are much more at risk. Thirdly, the overall weightage due to the number of people from 50-60, 60-80 and 80+ years old are much more than the less than 15 group.

Beds

The correlation coefficient for Beds with the number of COVID-19 deaths is observed as 1797.29. This is far larger than any of the other coefficients, which suggests that the number of beds has a more important role to play than any of the other parameters. Looking at our earlier hypothesis, we observe that the data is in stark contrast to that, as we expected a negative correlation, but we got a positive correlation, and a very strong one at that. This suggests that our initial hypothesis and reasoning is flawed, We now postulate an alternative hypothesis, which could serve to explain the data observed.

Firstly, it could be argued that the number of beds available is representative of the number of cases in an area. The reasoning is as follows. The number of beds in any area have increased since COVID-19 through the means of temporary shelters and quarantine zones, due to the sheer magnitude of the number of cases. The number of beds thus created would be dependent on the number of COVID-19 cases in an area. More COVID-19 cases mean more beds. This leads to a positive correlation between cases and beds. And where there are more cases, there are more deaths, thus explaining our positive correlation. Secondly, continuing the narrative of shortage of beds, it is also possible that cases were transported to the locations which have a greater number of beds available. Thus, even though the cases may not have originated from a particular location, they are using the beds of this location. Furthermore, these cases are most likely the serious ones for them to have been transported, which also means a higher likelihood of death. In this manner, the locations which have more beds can also be expected to have a higher number of COVID-19 deaths.

Democrats

The correlation coefficient for Democratic Percent with the number of COVID-19 deaths is observed as 9.32. This indicates a positive correlation between representation of Democrats, and the number of COVID-19 deaths. i.e., the more Democratic a State is, the more COVID-19 deaths are present. This is also in contrast to what our expected hypothesis was, which was a negative correlation as Democrats generally took more precautions against COVID-19.

Let us try to postulate why this may be so. If we look at COVID-19 deaths, most of them have been taking place in COVID-19 hotspots / clusters, due to the rampant spread of the COVID-

19 in these areas. These are mostly concentrated on urban or metropolitan areas. Coincidentally that is also where the majority of the people are Democrats, while Republicans have more of a following in rural areas. If that is the case, due to the higher rampancy of COVID-19 in these areas, it could explain the positive correlation in-spite of the extra precautions taken by Democrats. Simply put, COVID-19 clusters have more democrats, leading to more deaths amongst them.

Education

The correlation coefficient for Education with the number of COVID-19 deaths is observed as 183.7. This is the second highest in order when compared to the correlation coefficients of all other causal factors considered in this study. This observation is in stark contrast to the expected negative correlation. The more educated people were expected to take better precautions and avoid contraction and spread of the virus and thereby avoid deaths due to it. This contrary observation can be reasoned out in a couple of ways. Firstly, educated people stand the better chance of being employed especially in the situation of a pandemic and thereby constitute more of the working population. Therefore, they have more chances of contracting the virus in general when compared to the uneducated people who are more likely to be unemployed and thereby will be less exposed to the virus. Secondly because of the same skewed nature of employment and other opportunities the degree to which one has to travel varies with one's education. Thus this along with the other factors which arise with employment that cause risk puts the educated people in the danger zone of higher deaths. Thus the data corresponding to this is a better reflection of the effects of employment and the correlation for that is observed as expected. To have a better understanding as to how meticulously educated people follow precautions and how that leads to lesser deaths we need to compare statistics of employed people with similar features of crowding, contact, etc with the only difference being the education levels.

Income

The correlation coefficient for Income with the number of COVID-19 deaths is observed as 0.002. This is amongst the least in order when compared to the correlation coefficients of the other causal factors considered in this study. This implies that the data for other causal factors better fit this problem when compared to the relationship the income factor holds with the outcome of deaths due to COVID-19. A negative correlation was expected in the relationship between income and COVID-19 deaths as healthcare in general is very expensive and the people with more money muscle were expected to afford better healthcare and thereby reduce their chances of death. However the order of the correlation coefficient is too less to conclude whether this expected relationship is incorrect despite the sign of the same. It only implies that the other factors were more important when compared to income as a causal factor for COVID-19 deaths. A small counter reasoning that could be given as to why a positive correlation was observed is that older people are expected to have better incomes and savings as they both grow with experience and age. Thus, despite having the muscle to afford better healthcare they seem to be fundamentally at risk to COVID-19 biologically thus nullifying the benefit of financial muscle.

Population

The observed correlation between population in each state and the respective number of COVID-19 deaths is 0.0008. This is of a very small order as compared to the coefficients of the other parameters. This suggests that there is not much of an effect of the state's population on the number of COVID-19 deaths. Earlier we expected there to be a strong positive correlation, as the larger the population, the more will be the expected number of deaths as the virus can spread to more numbers of people. This is not what is observed here. To understand why this is, we need to take a closer look at our parameters. What we are considering here is the total population of the State. This neither represents how densely populated it is, neither does it show the presence of COVID-19 clusters. A statistic related to population which could be of more significance is state population density which refers to the total population of the state divided by the area of the state. This could quantify how dense populations are in a better manner and it is this clustering which fundamentally leads to higher human contact which could increase the number of infections and thereby deaths. The presence of clusters such as those in urban areas would agree with our earlier hypothesis of a positive correlation, while more spread-out areas like in rural places could have a lesser spread of the virus. However, such concentration areas /clusters are not captured effectively by the data here. Hence, we see that the state population data does not have much effect, as in some senses, it is random data without much useful information.

Temperature

The observed correlation between temperature and the number of COVID-19 deaths is -1.45. This indicates an inverse correlation between the number of COVID-19 deaths and the average temperature of the environment. The negative correlation matches with our expectations of temperature, with the main two reasons for this being as follows. Firstly, studies have found that people in places with warmer temperatures tend to be more resistant to the viruses as compared to people in colder temperatures. This is as people in warmer places tend to have higher levels of immunity on average. Secondly, recent studies into COVID-19 have found that the virus propagates better in places with lower temperatures, which explains the higher number of deaths at lower temperatures. An additional point to note is that temperature is particularly well reflected with the states data we have chosen, as these are geographically separate regions, with different locations and topography, but which within themselves have consistent/similar temperatures.

Conclusions

As stated earlier, our objective is to identify the dependence of deaths due to COVID-19 on various factors. The factors that we had chosen in our analysis were Median Age, number of Hospital Beds, Political Affiliation (specifically percentage of Democrats), Education (represented at the Bachelors level), Income, Population and Temperature. The effect was analyzed using the data of the various states of the USA for comparison, as this is easily available and is also consistent.

For the analysis, we used linear regression which basically tries to predict the number of COVID-19 deaths as a function of the other variables. The corresponding coefficients represent the weights of the various variables on the number of COVID-19 deaths. They can also be said to represent the correlation between the factor and the number of COVID-19 deaths. Using this, we can infer relations between the two. After applying linear regression, we obtained a set of coefficients which suggested that some variables had more importance than others, and also that some variables may not have had any effect at all. Specifically, we found Age, Beds, Democratic Affiliation and Education to have a positive correlation, while temperature has a negative correlation. Income and Population have almost no effect.

This data can be useful as using this, specific precautions can be taken to combat the spread of COVID-19. For example, if we know that age has a positive correlation, that would suggest that higher age groups are more at risk from COVID-19 and hence need to be more protected. More complicated strategies can also be developed.

One thing to note here is that our analysis is in some senses very limited, and can’t be taken as the absolute truth, Further analysis may be needed. For example, it could be that some results are only applicable in the case of the USA. This methodology can also be extended for more sophisticated analysis by using broader data sets, considering more factors, or using more appropriate fitting techniques. All this to help us better understand the virus.

References

- https://www.thehindu.com/news/national/analysis-how-pandemics-have-changed-the-world/article31345176.ece
- https://indianfolk.com/virus-shook-world/
- https://voxeu.org/article/economic-consequences-covid-19-multi-country-analysis
- https://en.wikipedia.org/wiki/Linear_regression
- https://machinelearningmastery.com/linear-regression-for-machine-learning/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7500589/
- https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html
- https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems
- https://link.springer.com/article/10.1007/s40808-020-00984-7
- https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-020-01781-w

Sources of Data:

1. Median Age:
   https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_median_age
2. Democrats per State:
   https://www.pewforum.org/religious-landscape-study/compare/party-affiliation/by/state/
3. Percentage of Bachelor Level Educated people per state:
   https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_educational_attainment
4. Number of beds per 1000:
   https://www.kff.org/other/state-indicator/beds-by-ownership/
5. Income:
   https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_GDP
6. Temperature:
   https://www.currentresults.com/Weather/US/average-annual-state-temperatures.php
7. Population:
   https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_population_density
8. COVID-19 Deaths:
   https://www.worldometers.info/coronavirus/country/us/