Investigating Predictors with Regards to Daily COVID-19 Related Fatalities[*]

A predictive model examining new COVID-19 related fatalities in North America

Adrian Wong

27 April 2021

Abstract

COVID-19 related fatalities are often associated with particular factors, such as government measures and hospital capacity. However, other factors are likely to influence the number of daily fatalities as well. Within this paper, a multiple regression model was proposed to identify key predictors with regards to daily COVID-19 related fatalities. Factors such as new cases per million, COVID reproduction rate, stringency index, population density, diabetes prevalence, number of female and male smokers, and number of hospital beds per thousand were factors associated with daily COVID-related deaths. By analyzing this data, results from this multiple regression model may help to predict and therefore reduce the number of COVID-related fatalities within North America.

1 Introduction

Data that was collected with regards to the Coronavirus from 2020 until 2021 can be effectively used to illustrate the progress made by countries to combat the global pandemic. According to *Our World In Data*, only if we end the pandemic everywhere can we end the pandemic anywhere. This is done by assessing the spread of COVID-19, its impact, our progress against it, and the efficacy of various measures (Max Roser and Hasell 2020a). As is the case, a predictive model that would help illustrate the connection between various COVID and location-based factors could potentially aid in the reduction of COVID-19 related fatalities.

This report is organized in two main sections: the data overview and the data analysis. Within the data overview, the source of the data will be discussed, as well as the variables within the data set and the methodology of analysis. Subsequently, the data analysis will explore several factors in relation to each location in North America, such as the COVID-19 reproduction rate over time, the government response stringency, hospital beds per 1000 people, new and total deaths from COVID-19, and population data. Moreover, each feature will be analysed in a correlation matrix to identify pairwise relationships. From this higher-level analysis, a multiple regression model will illustrate the predictor variables that have significant influence on new deaths due to COVID-19 and will arguably illustrate key factors to consider when examining COVID-19 related fatalities within this data set. It is in hopes that with this model, clear findings with regards to reducing the fatality of COVID-19 within North America will be found. Further discussion of key findings, limitations, and next steps will be presented in conclusion.

[*] Code and data are available at: https://github.com/A1001949/covid_predict.git

2 Data Overview

2.1 Source

This data set was collected by Our World In Data in order to study the global situation with regards to the Coronavirus Pandemic. With this data, the organization seeks to answer questions such as *Is it possible to make progress against the pandemic?* and if so, *How can we make progress against the pandemic?* At an organizational level, their goals are to develop a *comprehensive perspective on global living conditions and the earth's environment* in order to see *powerful changes that reshape our world* (Max Roser and Hasell 2020a).

Rows from this data set were collected from various sources. In particular, confirmed cases and deaths were sourced from the COVID-19 Data Repository by the CSSE at Johns Hopkins University ("COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University" 2021). Data with regards to hospitalizations and ICU admissions came from the European Centre for Disease Prevention and Control ("Data on Hospital and ICU Admission Rates and Current Occupancy for COVID-19" 2021), the government of the UK ("GOV.UK Coronavirus (COVID-19) in the UK" 2021), the Department of Health & Human Services ("Covid19 Reported Patient Impact and Hospital Capacity State Timeseries" 2021), and the COVID-19 Tracker ("COVID-19 Tracker Canada" 2021). A limitation for this data set was that information with regards to hospitalizations for other countries were unable to be collected as there is no such aggregated data, nor does Our World in Data have the capacity to store such data. COVID-19 testing and vaccinations are a relatively new topic in terms of data gathering and testing data is updated twice a week by the OWID team. Finally, other variables within this data set are sourced from other sources such as United Nations, World Bank, Global Burden of Disease, Blavatnik School of Government, etc. (Max Roser and Hasell 2020a).

Another limitation from this data set is that although it is ongoing, it is incomplete. Data sources are dated as early as January 21st, 2020, and data such as total deaths, vaccinations, and testing were simply non-existent at the time. Moreover, many of these data sources are updated on a weekly, biweekly, or most recent basis. As is the case, several data points do not accurately reflect the day-to-day changes that are necessary to build a perfectly accurate picture of the actual data in the world. For example, *population* represents the population of a given location in 2020 and in no way reflect day-to-day changes in the total population.

This data set is regularly maintained by Our World in Data and is updated daily. A changelog is posted on their public GitHub for users to keep track of major changes and for transparency's sake (Max Roser and Hasell 2020a).

2.2 Variables

Within this data set, there were **10014** responses and 59 features in total. These features included data with regards to the location of the observation, the date of which the observation was made, new and total case data, new and total death data, number of ICU patients and admissions, number of hospital patients and admissions, new and total tests, vaccinations, and other location-based data such as population, GDP per capita, age distribution, handwashing facilities, and life expectancy.

Finally, 13 features were chosen to be a part of the regression model including:

- `total_deaths` : Total deaths attributed to COVID-19
- `new_deaths` : New deaths attributed to COVID-19 (Day)
- `total_cases_per_million` : Total confirmed cases of COVID-19 per 1,000,000 people
- `new_cases_per_million` : New confirmed cases of COVID-19 per 1,000,000 people
- `new_cases_smoothed_per_million` : New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people

- `reproduction_rate` : Real-time estimate of the effective reproduction rate (R) of COVID-19
- `stringency_index` : Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)
- `population` : Population in 2020
- `population_density` : Number of people divided by land area, measured in square kilometers, most recent year available
- `diabetes_prevalence` : Diabetes prevalence (% of population aged 20 to 79) in 2017
- `female_smokers` : Share of women who smoke, most recent year available
- `male_smokers` : Share of men who smoke, most recent year available
- `hospital_beds_per_thousand` : Hospital beds per 1,000 people, most recent year available since 2010

Factors were not chosen for the multiple regression model if they were non-numerical (e.g. Continent), missing more than half of the column data, or had high pairwise correlation in order to account for multicollinearity.

2.3 Methodology

To analyze this data set, `R` statistical programming language (R Core Team 2020) was used to import the respective data set from the Our World in Data website through stable URLs for reproducability (Max Roser and Hasell 2020a). The `here` package was used to construct a relative file path to the data set within the repository (Müller 2020). For data manipulation, exploration, and visualization, the `tidyverse` package was used as well (Wickham et al. 2019). In particular, graphs were created using `ggplot2` (Wickham 2016) and a table was constructed using `kable` (from the `knitr` package (Xie 2020)). `caret` was used in order to identify highly correlated factors to remove (Kuhn 2020) and `corrplot` was used to construct a correlation matrix plot (Wei and Simko 2017). The `broom` package was used for tidying the model's statistical findings in a tibble (Robinson, Hayes, and Couch 2021). The `scales` package was used to organize p-values into proper notation (Wickham 2018). Finally, `janitor` was used to clean the data (Firke 2021).

Once the data was imported, an `R` script were used to clean the data. This was done by first formatting the column names such that capitalization and spaces were removed in favor of underscores. Moreover, given that this analysis was focused on predicting new deaths in North America using several COVID-19 factors, data from other continents were removed. The original data set contained 82,837 observations and 10,014 observations remained after data cleaning.

After descriptive analysis, several features were removed in order to build the multiple regression model. Non-numeric information was removed from the dataframe, such as location-based information (`iso_code`, `continent`, `location`, `date`, `test_unit`).

Next, 26 features that had more than 50% missing data was removed as well (see limitations in the Variables section above; `icu_patients`, `icu_patients_per_million`, `hosp_patients`, `hosp_patients_per_million`, `weekly_icu_admissions`, `weekly_icu_admissions_per_million`, `weekly_hosp_admissions`, `new_tests`, `weekly_hosp_admissions_per_million`, `total_tests`, `total_tests_per_thousand`, `new_tests_per_thousand`, `new_tests_smoothed`, `new_tests_smoothed_per_thousand`, `positive_rate`, `tests_per_case`, `total_vaccinations`, `people_vaccinated`, `people_fully_vaccinated`, `new_vaccinations`, `new_vaccinations_smoothed`, `total_vaccinations_per_hundred`, `people_vaccinated_per_hundred`, `people_fully_vaccinated_per_hundred`, `new_vaccinations_smoothed_per_million`, `people_fully_vaccinated_per_hundred`). `new_vaccinations_smoothed_per_million`, `extreme_poverty`).

Finally, correlational analysis was performed to identify highly correlated factors. To address multicollinearity, 15 features were removed as well (`gdp_per_capita`, `new_deaths_smoothed_per_million`, `human_development_index`, `new_cases_smoothed`, `new_deaths_smoothed`, `total_deaths_per_million`, `new_cases`, `cardiovasc_death_rate`, `total_cases`, `new_deaths_per_million`, `handwashing_facilities`, `aged_65_older`, `life_expectancy`, `aged_70_older`, `median_age`).

3 Data Analysis

This data set includes data with regards to the current global COVID-19 pandemic on a daily basis. Among the features being recorded per country within North America, the COVID-19 reproduction rate gives a real-time estimate of the amount of other individuals a patient with COVID-19 can naturally infect on average. As is the case, this variable is an especially interesting variable to note, as it could potentially account for the amount of new cases within a given location ("Coronavirus: What Is the r Number and How Is It Calculated?" 2021).

3.1 COVID-19 Reproduction Rate

In the beginning of the global pandemic, the reproduction rate was highest in the United States and Canada, averaging around 3 - this implies that an American or Canadian citizen with COVID-19 was likely to infect 3 others during the onset of the pandemic when data was initially being collected. This makes sense given that the reproduction and implications of COVID-19 was too early to be known at the time. Currently, the locations with the highest reproduction rate is Trinidad and Tobago, while the reproduction rate of COVID-19 in Grenada is approaching 0 (see figure 1).

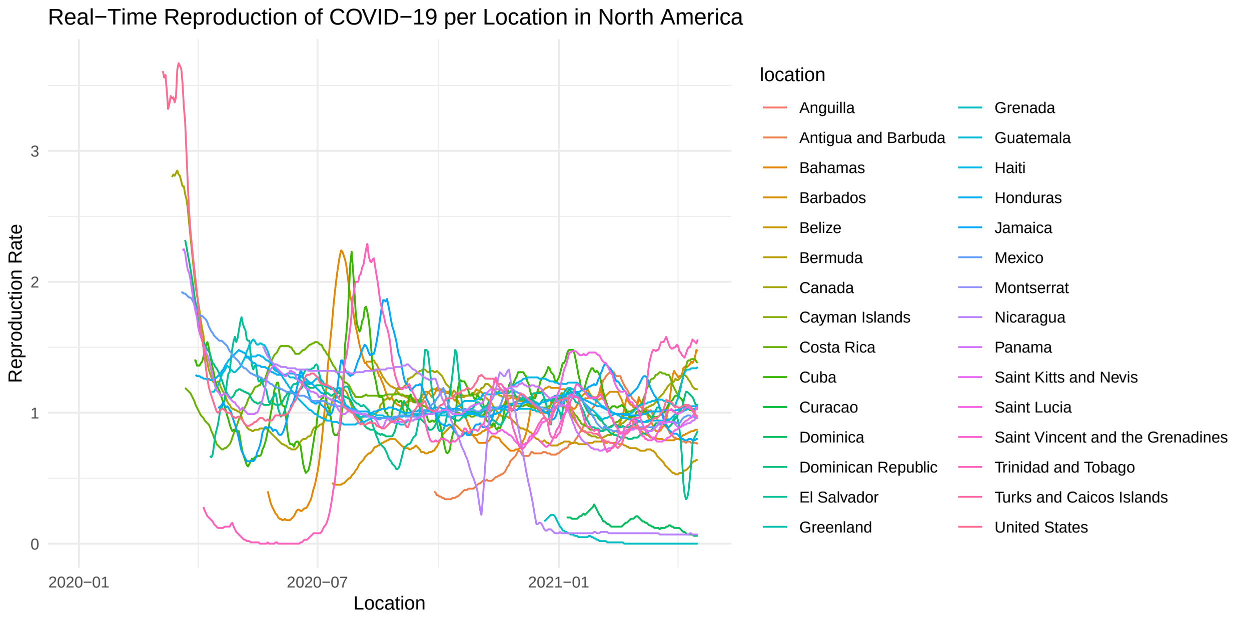


Figure 1: COVID-19 reproduction rate holds relatively steady but differs by location

3.2 Government Response Stringency

Naturally, the government within each location approaches COVID-19 measures differently. Government response stringency is a composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 where the strictest response is 100.

If the average stringency index per location is analyzed, we see that Honduras has taken the strictest measures against COVID. Currently, Honduras is rated as COVID-19 level 4 (very high) and even fully vaccinated travelers are recommended not to travel there ("COVID-19 in Honduras" 2021). As for Canada and the United States, the Government response stringency index hovers around 63/100 where the average stringency index of all countries across the past year is 63.5169842. Thus, the average stringency index across North America was relatively strict (see figure 2).

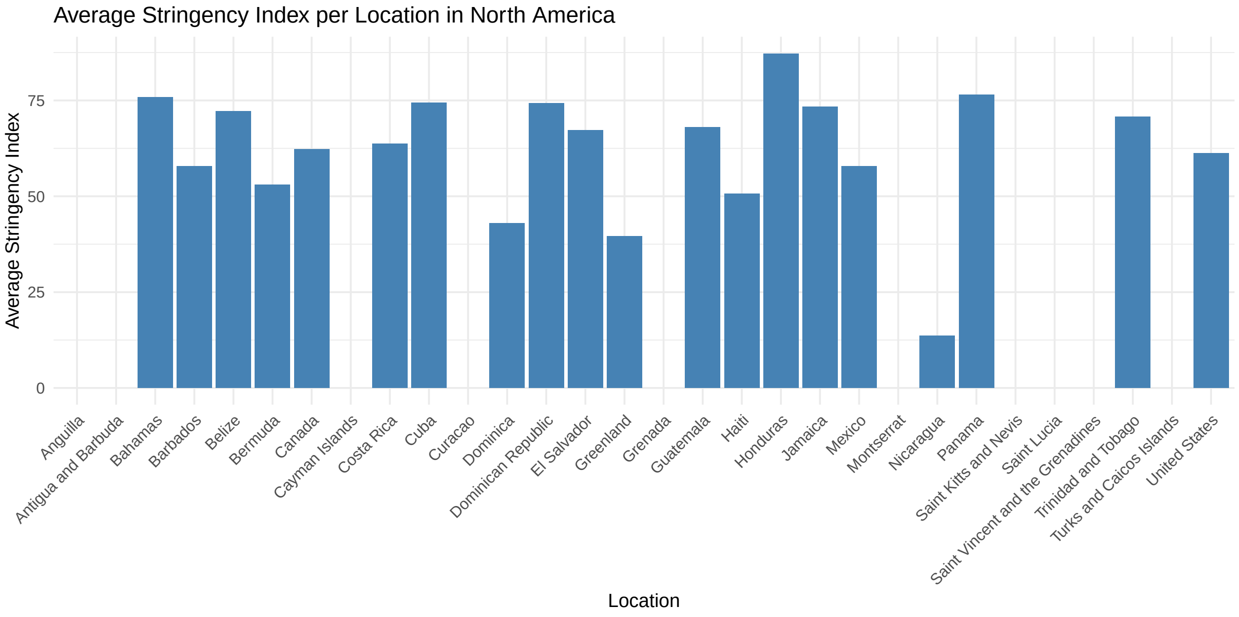


Figure 2: Government Stringency Responses have been on average more strict than not (>50)

3.3 Hospital Beds per 1,000

Nonetheless, the amount of hospital beds vary according to the location. The data set collects the number of hospital beds per 1,000 people based on the most recent year ever since 2010. As is the case, locations that have more than 1000 hospital beds have more hospital beds per people and locations that have less than 1000 have less hospital beds per people.

According to figure 3, Barbados is the location with the most hospital beds per 1,000 people, where Guatemala, Haiti, and Honduras have among the least. The United States and Canada have approximately 1,000 hospital beds per 1,000 people.

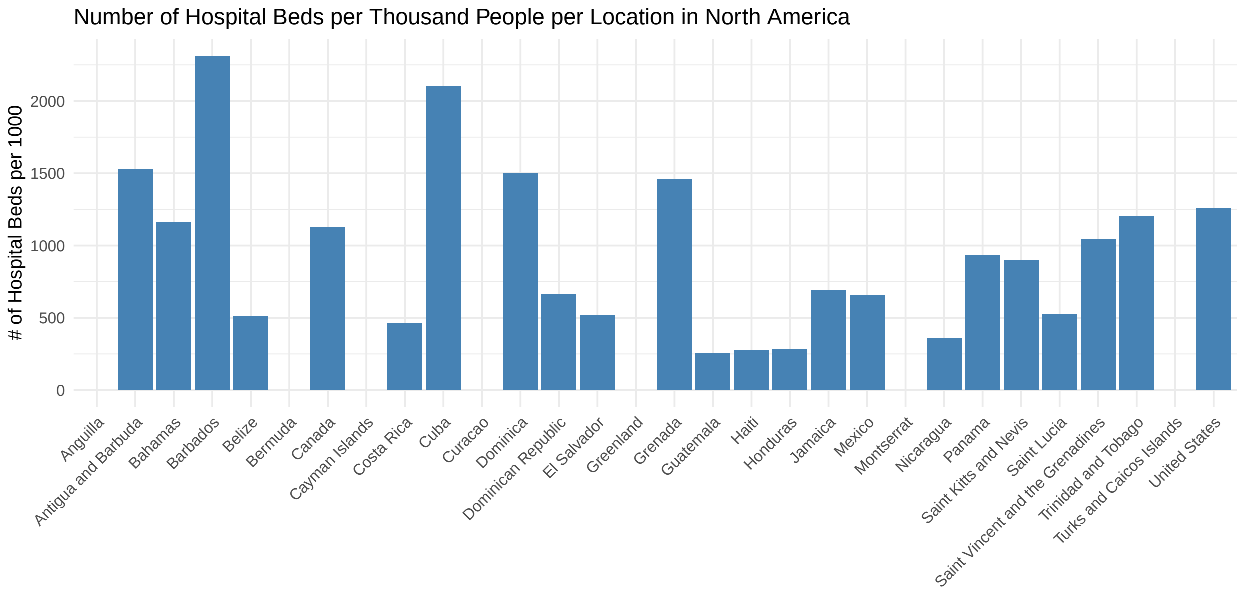


Figure 3: The number of hospital beds available vary per location

3.4 New Deaths

With more hospital beds and higher stringency rates, one hypothesis could potentially be that the amount of new COVID-19 related fatalities will be less in those particular locations. However, this does not necessarily seem to be the case.

Figure 4 illustrates clearly that the countries with the highest amount of COVID-19 related deaths per day were the United States, Canada, and Mexico. Within those three locations, new fatalities seemed to peak around New Years of 2021, with the United States having roughly 4000 new deaths per day.

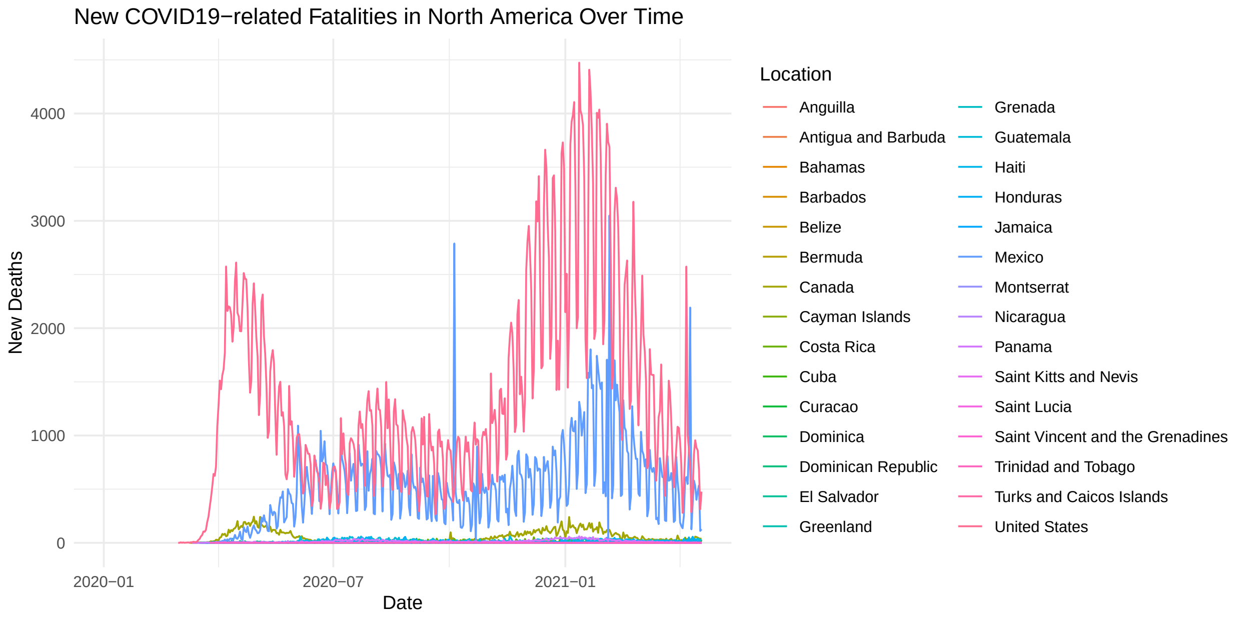


Figure 4: COVID19-related fatalities spiked near January 2021

3.5 Total Deaths

When visualizing the total deaths over time, this pattern seems to be more clear - United States, Mexico, and Canada were the locations were the countries with the highest deaths in North America due to COVID-19 (see figure 5). This was regardless of the fact that Canada and the United States had a government response stringency index close to the average of North America and had enough hospital beds at roughly a 1:1 ratio among 1000 people. Thus, more variables are at play when considering COVID-19 related deaths in North America.

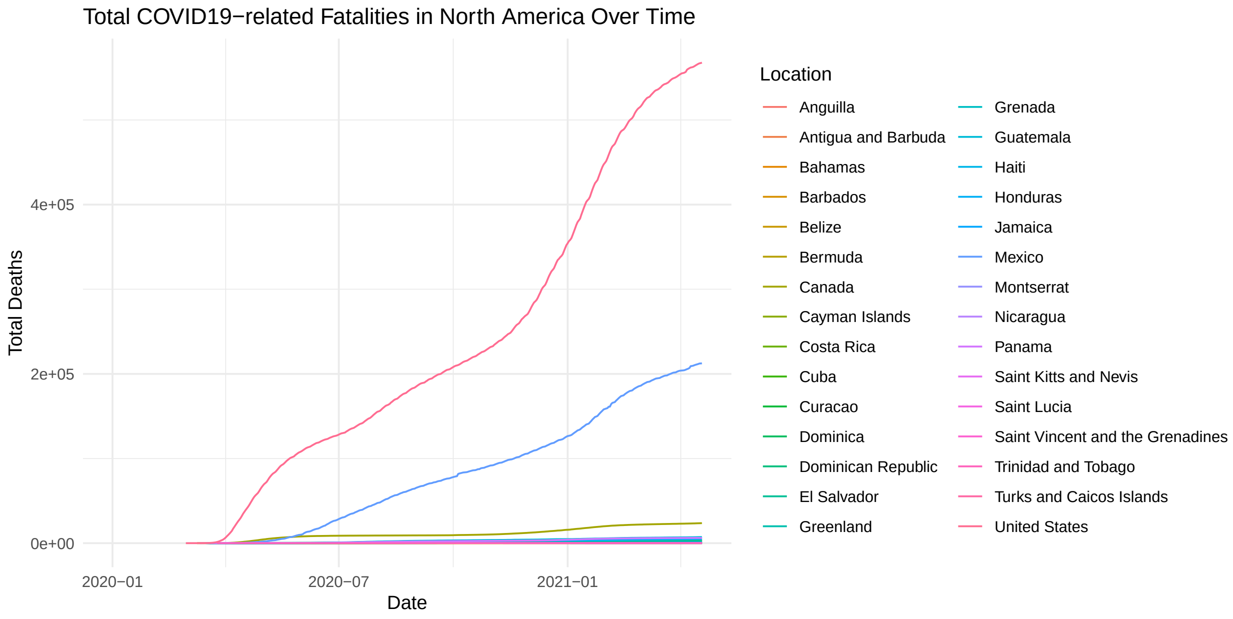
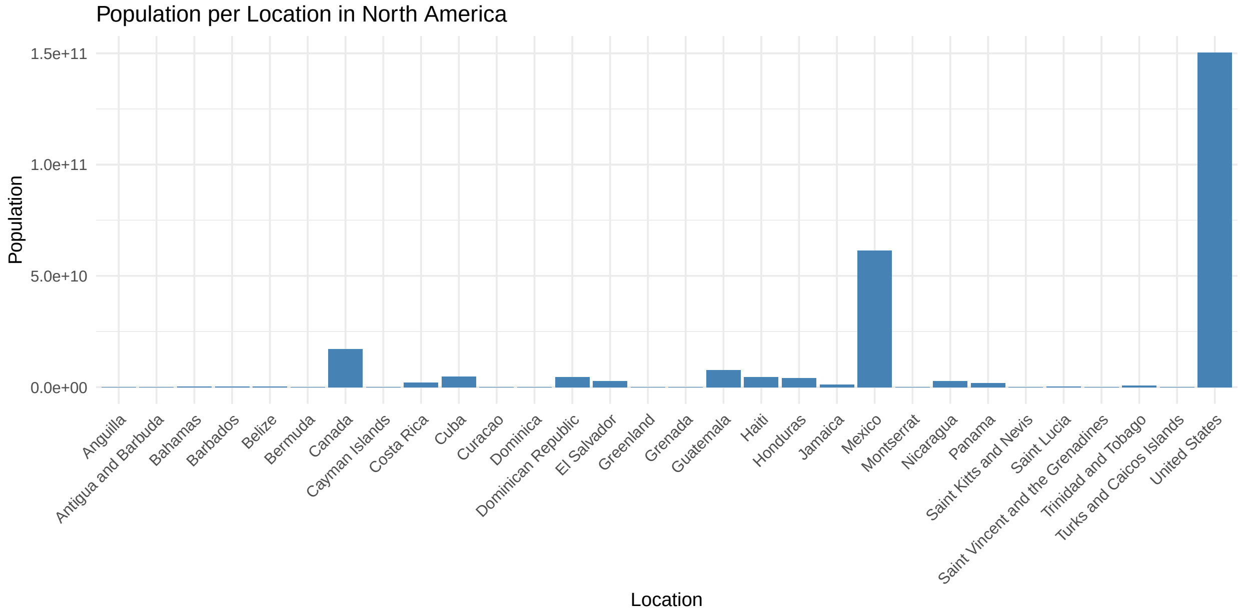


Figure 5: Total COVID-19 Related Deaths Increased Over Time

3.6 Considering Population

A significant variable that must be considered is the population of each location. Notably, the US, Canada, and Mexico have the highest population in North America. This seems to correspond to the high amount of COVID-19 related fatalities in these locations in comparison to the rest of North America (see figure **??**).



3.7 Factor Correlations

Given that population seems to be related to COVID-19 related fatalities, it is important to also identify other factors that are intercorrelated as well. This correlation analysis can help identify factors of interest and address multicollinearity if two predictor variables have high pairwise correlation. Predictors with high correlation should not be mapped to the same linear model.

Several variables were not included in this correlational analysis. For example, the location, test_unit, and dates have no explanatory power in the linear model and were not appropriate for correlational analysis due to non-numerical data types. Moreover, features that had more than half of their data missing were removed from this analysis as well. This is a significant limitation to highlight given that vaccinations and testing were relatively new datapoints collected recently. As a result, many observations related to vaccinations and testing were not included in the correlational analysis, even though they may have been significant predictors of new COVID related fatalities. In the end, 28 variables remained for correlational analysis.

Figure 6 illustrates the relationship between these factors, including *new_deaths* which will be the outcome variable of the proposed multiple regression model. Notably, several variables are highly correlated to one another due to calculations. For example, *median_age*, *aged_65_older*, and *aged_70_older* are naturally highly correlated to each other. Other correlations were insightful though expected, such as the correlation between *life_expectancy* and *human_development_index.* Some correlations were insightful as well as surprising (*handwashing_facilities* and *cardiovasc_death_rate* were negatively correlated by -0.97).

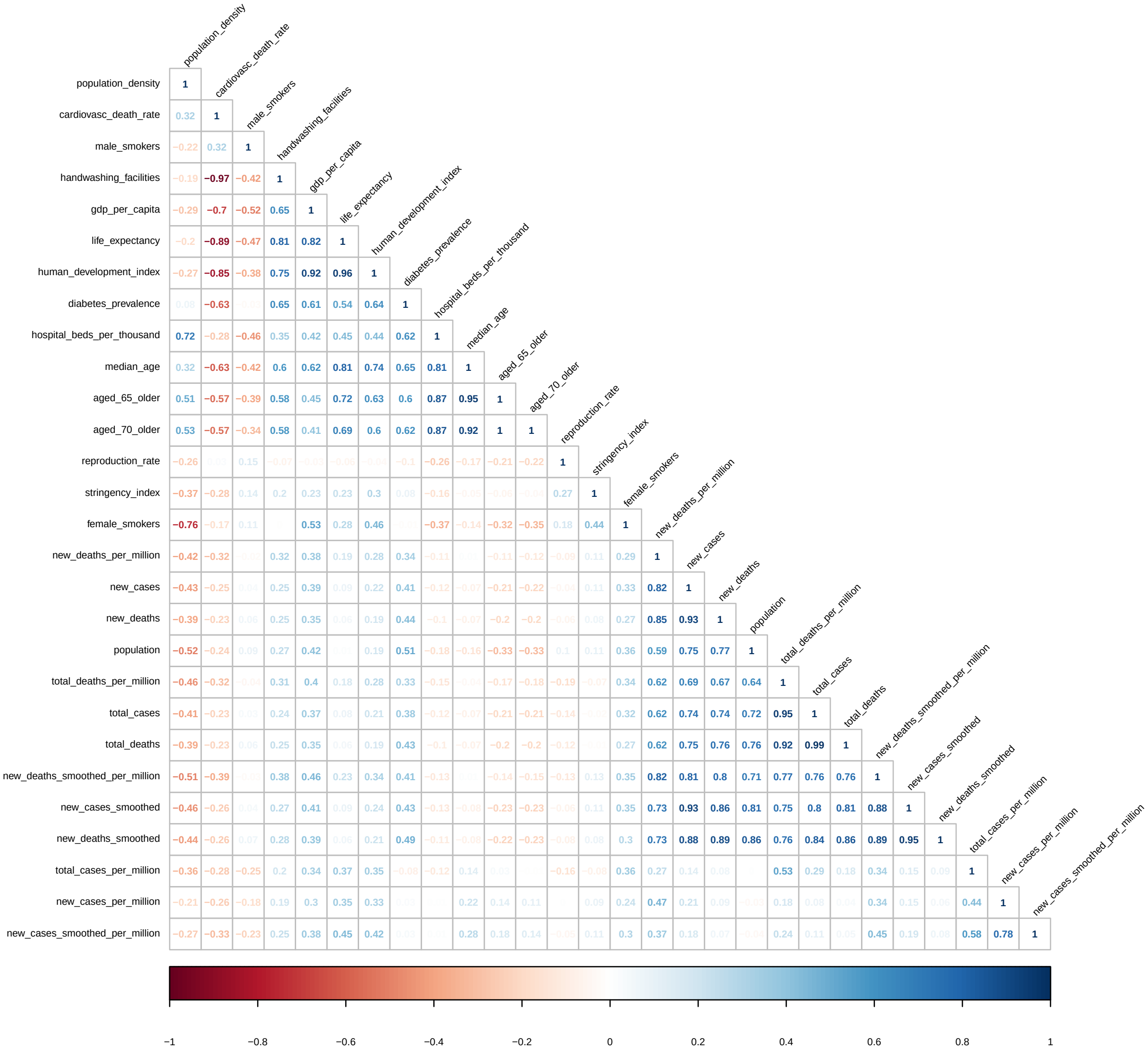


Figure 6: Many factors within the data set have high pairwise correlations

new_deaths were highly correlated to multiple factors including *population* (0.77), *total_cases* (0.74), and *new_cases* (0.93). This feature was only moderately correlated to factors such as *population_density* (-0.39), *gdp_per_capita* (0.35), and *diabetes_prevalence* (0.44). Surprisingly, the hypothesis that *stringency_index* and *reproduction_rate* were highly correlated with *new_deaths* was unfounded, as their correlations were merely 0.08 and -0.06 respectively.

3.8 Linear Model

Given these correlations, is it possible that various factors can predict new deaths due to COVID-19 when considered in tandem to one another?

In order to answer this research question, a multiple regression model was created in RStudio. Multiple linear regression was chosen because the research question entails a single dependent variable (*new_deaths*) and wants to evaluate the relationship between such a variable with several predictor variables. Moreover, a multiple regression model has the ability to take factors into account that may not seem very important. Finally, unlike other models that require other data types such as logistic regression, multiple regression only requires continuous values for predictor variables and the dependent variable. As is the case, multiple regression model is a suitable model in order to determine whether continuous factors can accurately predict new deaths related to COVID within this data set.

Multiple regression models come with the following assumptions ("Assumptions of Multiple Linear Regression," n.d.):

1. No Multicollinearity — Multiple regression assumes that the independent variables are not highly correlated with each other.
2. Linear Relationship – There must be a linear relationship between the independent variables and the outcome variable.
3. Multivariate Normality – Multiple regression assumes that the residuals are normally distributed.
4. Homoscedasticity – This assumption states that the variance of error terms are similar across the values of the independent variables.

3.8.1 Lacking Multicollinearity

The previous correlational analysis was useful to prove a lack of multicollinearity between predictors. Moreover, the `findCorrelation()` function from the `caret` package was useful to remove factors with high multicollinearity (that is, factors with correlation coefficients higher than +0.8 or -0.8) (Kuhn 2020).

3.8.2 Linear Relationship

By checking the Residuals vs Fitted plot (see figure 7), we can verify our assumption that the independent variables have a linear relationship to the dependent variable. Within this plot, we see that the red line is approximately horizontal near 0, suggesting that linearity can be assumed between the predictors and the outcome variable.

3.8.3 Multivariate Normality

To verify multivariate normality, a Normal Q-Q plot can be checked to see how well the points follow the dashed line (see figure 7). In this model, we can assume normality given that most points follow the dashed line, but notice heavy tails within the plot.

3.8.4 Homoscedasticity

A Scale-Location plot would be useful to verify homoscedasticity (see figure 7). In this plot, a horizontal line with points spread evenly across the plot would indicate homoscedasticity.

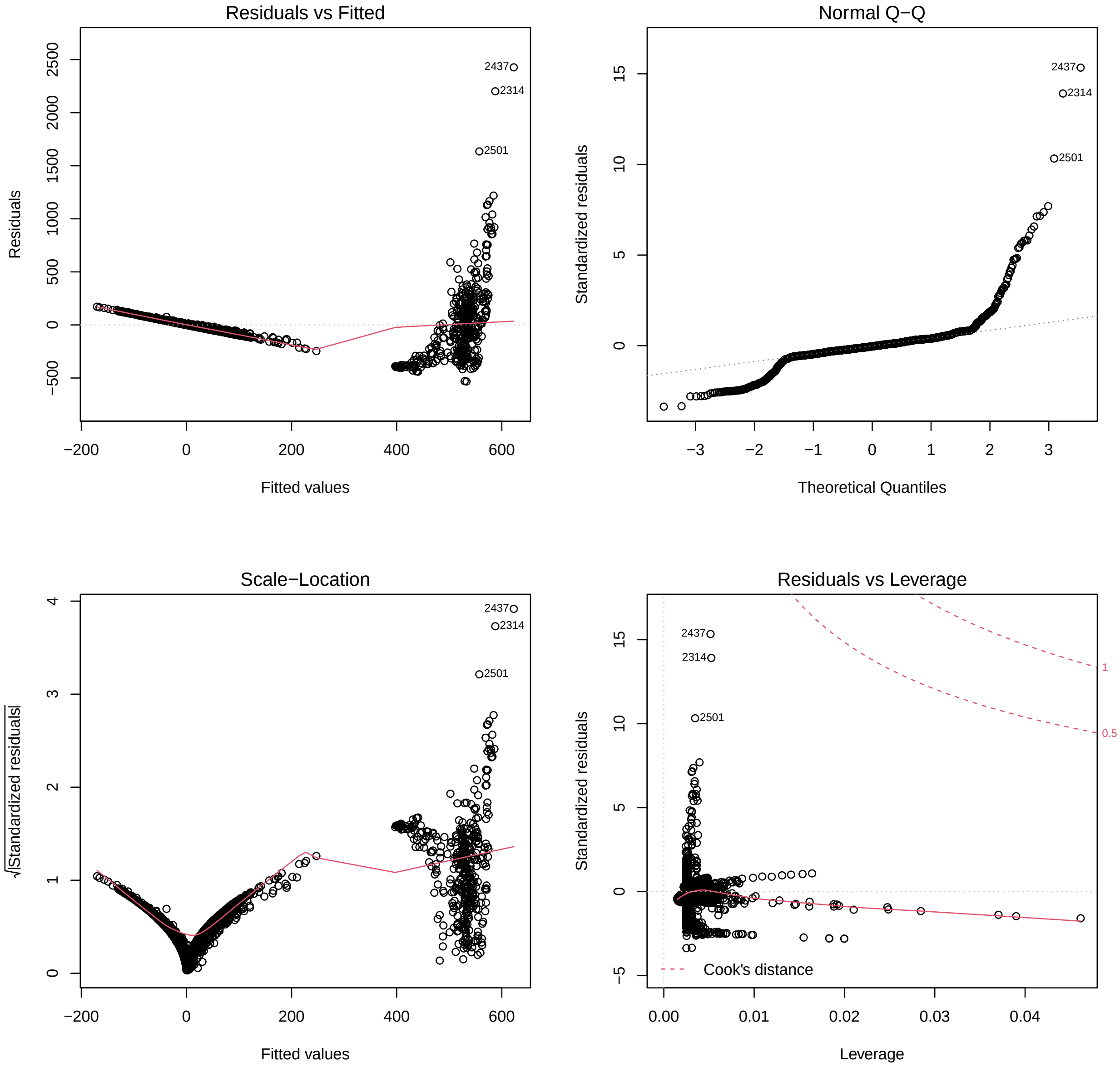


Figure 7: Senior patients have significantly more fatal cases than young patients

4 Results

Note that in table 1, all eight predictors had significant t-values where their p-values were <0.001. After creating the multiple regression model, the equation for the proposed model is as follows:

$$y = 0.47x_1 - 135.31x_2 - 0.75x_3 + 2.81x_4 + 221.41x_5 + 87.65x_6 - 50.68x_7 - 516.96x_8 - 1183.07$$

where y represents *new_deaths* (the outcome variable), x_1 represents *new_cases_per_million*, x_2 represents *reproduction_rate*, x_3 represents *stringency_index*, x_4 represents *population_density*, x_5

Table 1: Multiple Linear Regression Against New Deaths due to COVID-19

Predictor	Estimate	SE	t-value	p-value
(Intercept)	-1183.0655729	49.1575358	-24.066820	<0.001
new_cases_per_million	0.4659052	0.0474691	9.814925	<0.001
reproduction_rate	-135.3117136	15.4808550	-8.740584	<0.001
stringency_index	-0.7518290	0.2050664	-3.666272	<0.001
population_density	2.8134686	0.1063814	26.446985	<0.001
diabetes_prevalence	221.4088935	4.9922602	44.350432	<0.001
female_smokers	87.6470820	3.5192432	24.905094	<0.001
male_smokers	-50.6767433	1.6415567	-30.871150	<0.001
hospital_beds_per_thousand	-516.9591351	14.5758306	-35.466873	<0.001

represents *diabetes_prevalence*, x_6 represents *female_smokers*, x_7 represents *male_smokers*, and x_8 represents *hospital_beds_per_thousand.*

In the model, the R-Squared value was 0.604 which means that the predictors have an explanatory power of 60% of the variance of *new_deaths.* Moreover, the p-value of the overall regression model was <0.001 so the model was statistically significant.

5 Discussion

5.1 Key Findings

Various predictor variables can account for approximately 60% of the variance in new deaths as a result of COVID-19. Assuming other factors were held constant:

- as new cases of COVID-19 per million increases, so does the mean number of new deaths by only 0.47 units.
- interestingly, the model also suggests that as the reproduction rate of COVID-19 increases, new fatalities decrease by -135.31 units.
- as was hypothesized previously, as stringency index increases, the mean number of new deaths decrease by -0.75 units.
- unlike the overall population, increase in the population density accounts for only a 2.81 unit increase in the mean number of new deaths.
- an interesting finding was that as the prevalence of diabetes increases, mean number of new deaths increase by 221.41 units.
- as the number of female smokers increase, so does the mean number of new deaths by 87.65 units.
- unlike female smokers, increasing the number of male smokers account for a -50.68 unit decrease in the mean number of new deaths.
- as is expected, increasing the number of hospital beds per thousand people decreases the mean number of new deaths by -516.96 units.

5.2 Limitations

Several limitations should be pointed out in this study. First, this multiple regression model indicates predictive correlation and not causation. It would be reckless to assume that factors such as the number of female smokers (and conversely, not the number of male smokers) are positively attributed to the number of COVID-19 related deaths. This is naturally the case with multiple regression models. Moreover, as temporal precedence was not established, one cannot assume causal inference from this model.

Additionally, it must be recognized that several factors were not included in the regression model for reasons listed previously. However, this does not necessarily imply that they were not important for predicting new deaths due to COVID-19. For example, vaccinations would hypothetically reduce the number of new COVID-19 cases and consequently reduce the number of new COVID-related fatalities as well. Moreover, many rows that had information that was not available and were removed in order to conduct correctional analysis and the multiple regression model.

Finally, this analysis only looks at data from North America. By focusing on locations with similar geography, we try to address cultural and geographical confounds that may influence the model. However, this may imply that the model is only applicable to North America and lacks external validity.

5.3 Next Steps

Future studies can take a more experimental approach to identify causal factors that may better predict new deaths due to COVID-19. Interesting findings suggested that government response stringency had low predictive power in reducing the amount of new COVID-related deaths as the measure is defined. Perhaps a future experiment could determine particular factors that could potentially influence government decisions in reducing COVID-19 related deaths more effectively.

Furthermore, a future study could build a multiple regression model using future data as it pertains to vaccinations and testing. To date, there is not enough data to make a significant model as vaccination and testing data is missing for most observations within this data set. As more data is gather around these two topics, a regression model could include these as important predictors.

Lastly, applying this study to a broader scale may improve the generalizability of this model to outside of North America. By applying data from other continents, the aim to reduce COVID-19 related deaths could be more comprehensive and therefore, more effective as a whole. It would also be rewarding to see if the multiple regression model holds for continents with other cultures and geographical locations.

Appendix

Reference

"Assumptions of Multiple Linear Regression." n.d. https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/.

"Coronavirus: What Is the r Number and How Is It Calculated?" 2021. https://www.bbc.com/news/health-52473523#:%C2%A0:text=A.

"COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University." 2021. https://github.com/CSSEGISandData/COVID-19.

"COVID-19 in Honduras." 2021. https://wwwnc.cdc.gov/travel/notices/covid-4/coronavirus-honduras.

"Covid19 Reported Patient Impact and Hospital Capacity State Timeseries." 2021. https://healthdata.gov/dataset/covid-19-reported-patient-impact-and-hospital-capacity-state-timeseries.

"COVID-19 Tracker Canada." 2021. https://covid19tracker.ca/.

"Data on Hospital and ICU Admission Rates and Current Occupancy for COVID-19." 2021. https://www.ecdc.europa.eu/en/publications-data/download-data-hospital-and-icu-admission-rates-and-current-occupancy-covid-19.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.* https://CRAN.R-project.org/package=janitor.

"GOV.UK Coronavirus (COVID-19) in the UK." 2021. https://coronavirus.data.gov.uk/details/healthcare.

Kuhn, Max. 2020. *Caret: Classification and Regression Training.* https://CRAN.R-project.org/package=caret.

Max Roser, Esteban Ortiz-Ospina, Hannah Ritchie, and Joe Hasell. 2020b. "Coronavirus Pandemic (COVID-19)." *Our World in Data.*

———. 2020a. "Coronavirus Pandemic (COVID-19)." *Our World in Data.*

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files.* https://CRAN.R-project.org/package=here.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2021. *Broom: Convert Statistical Objects into Tidy Tibbles.* https://CRAN.R-project.org/package=broom.

Wei, Taiyun, and Viliam Simko. 2017. *R Package "Corrplot": Visualization of a Correlation Matrix.* https://github.com/taiyun/corrplot.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2018. *Scales: Scale Functions for Visualization.* https://CRAN.R-project.org/package=scales.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2015. *Dynamic Documents with R and Knitr.* 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://yihui.org/knitr/.