STAT 5232 Final Project

Ali Turfah

5/17/2020

Overview

This project is an attempt to fit a model to predict the daily Coronavirus death rate at a country-level. Data is obtained through the Census.gov Global Information Gateway and the COVID-19 Data Repository at Johns Hopkins, which are used to compute the model country-level properties as well as the daily number of death, recovery, and infection incidents respectively. This report is written based on the data available on May 17. Overall the model is capable of capturing the trend in Coronavirus death rates for certain countries, but there are cases where its ability to accurately forecast is noticeably limited.

Data Preparation

The data sources used are the Census.gov Global Information Gateway and the COVID-19 Data Repository at Johns Hopkins. The Census.gov data is used for country-level information such as overall population size and the distribution of the population over different age groups. The Johns Hopkins COVID-19 Data repository has day-by-day reporting of the cumulative number of cases, recoveries, and deaths at a country and territory-level.

The data cleaning steps are as follows. The Census.gov data has duplicate entries for Guinea, which are omitted as these cannot be differentiated. Along the same lines, countries that do not line up between the two data sources are also not considered in the analysis. In addition, the five most recent days of data are ignored due to concerns about lags in reporting time and these updates being reflected in the files. Any territory/state level results (e.g. Australia and Canada) are aggregated to the overall country level. Finally, the cumulative case, recovery, and death values are transformed into the daily number of incidents.

For all remaining country-day level observations, days are considered relative to the first confirmed case of COVID-19. For example, if Canada's first confirmed case is on Jan. 1 and the United States's first case is Jan. 15, then these would be the day 1 entries for Canada and the U.S., respectively. Taking a nod from the Succeptible-Infected-Recovered (SIR) model, the number of Infected patients at day i is calculated as

$$Infected_i = \sum_{j<i} Confirmed_j - Recovered_j - Deaths_j$$

Where $Confirmed_j$, $Recovered_j$, and $Deaths_j$ correspond to the number of new confirmed cases, recovered cases, and reported deaths on day j. This gives a running total of the number of people who are currently infected, which is the possible pool of people who can die from the virus on day i. The death rate for a date i is then calculated as

$$DR_i = \frac{Deaths_i}{Infected_i}$$

Any countries that have not yet had any deaths are omitted from the analysis; these countries are either too early in the process to provide meaningful data or their data is not fully up-to-date. After these filtering steps, the data for 148 countries remain.

This final data set is split into training and evaluation subsets, where the evaluation subset contains the most recent 15 days of data. This is meant to assess the model's predictive ability and to see the kinds of trends it can pick up on.

Model Selection

I will consider one model and three sub-models in this analysis. The overall model of interest (DC-model) is a Quasi-Poisson GLM where the the link function is of the form

$$log\left(\mathbb{E}\big[Deaths_i\big]\right) = \beta_0 + \beta_1 D_i + \beta_2 C_i + \beta_3 D_i C_i + log(Infected_i)$$

Where D_i is the number of days since the first confirmed case, C_i is a categorical variable for the country of observation i, along with an interaction term between the two covariates. There is also an offset term for the log of the number of infected people; this allows the overall model to be fit for the death rate (DR_i defined above). The Quasi-poisson is chosen over the Poisson because with such a simple model the assumption that there are omitted covariates, which would lead to overdispersion, is almost certainly valid. The simplified models are the C-model which only considers a country-level categorical variable, the D-model which only considers the D_i variable, and the null model which only fits an intercept term.

As these are nested models, they can be compared with the Residual Deviance. This metric prefers the DC model to the other three, as shown below. Note that no model had a good residual deviance relative to the degrees of freedom, which suggests an overall inadequate fit of the data.

Table 1: Performance of different models on training data

Model Type	Residual Degrees of Freedom	Residual Deviance
DC	8181	43170.14
D	8475	147954.88
C	8329	77349.85
NULL	8476	148724.10

Results

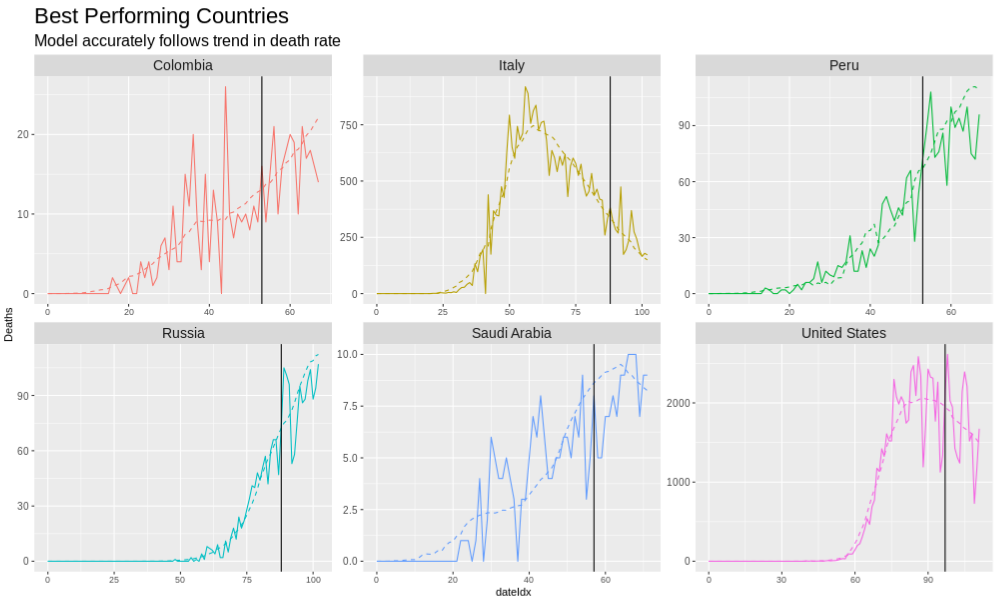
Model performance on the testing set is determined by the absolute percent error in the prediction, which is calculated as

$$err = \sum_{i=1}^{n} \frac{|true_i - predicted_i|}{true_i}$$

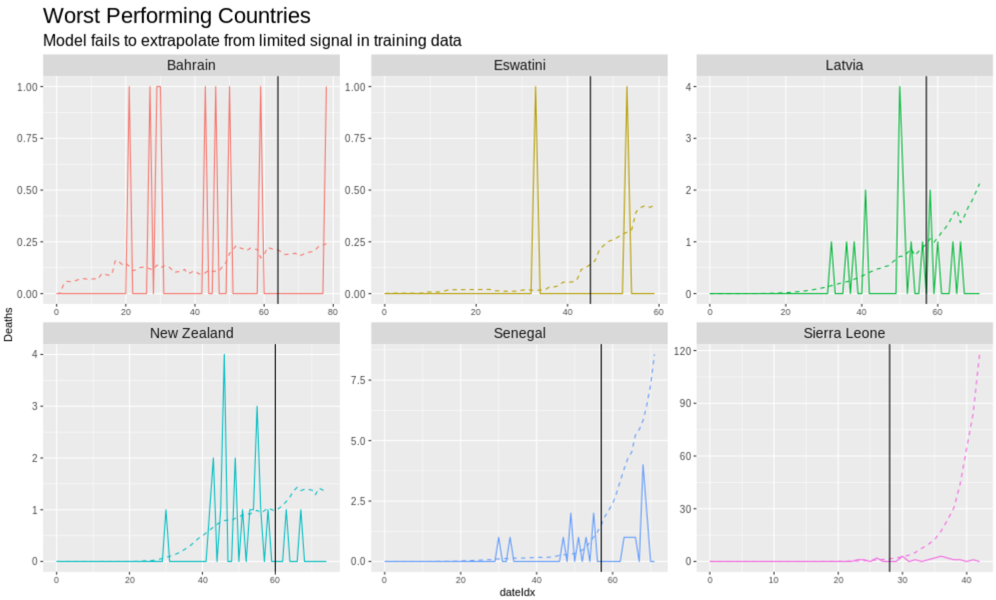
The main benefit of this metric is that errors are controlled to the scale of the value that is being predicted; over-predicting 100,000 deaths by 1,000 is much less egregious than over-predicting 100 deaths by that same amount. To get a sense of the types of trends the model captures, I will examine the overall performance of the model in three cases: the best performing countries, the worst performing countries, and the countries with the highest number of deaths.

In the plots below, the solid and dotted lines are the actual and predicted number of deaths respectively. The vertical black line is the threshold between the training and evaluation sets. A more detailed view of the evaluation set results is included in the final section.

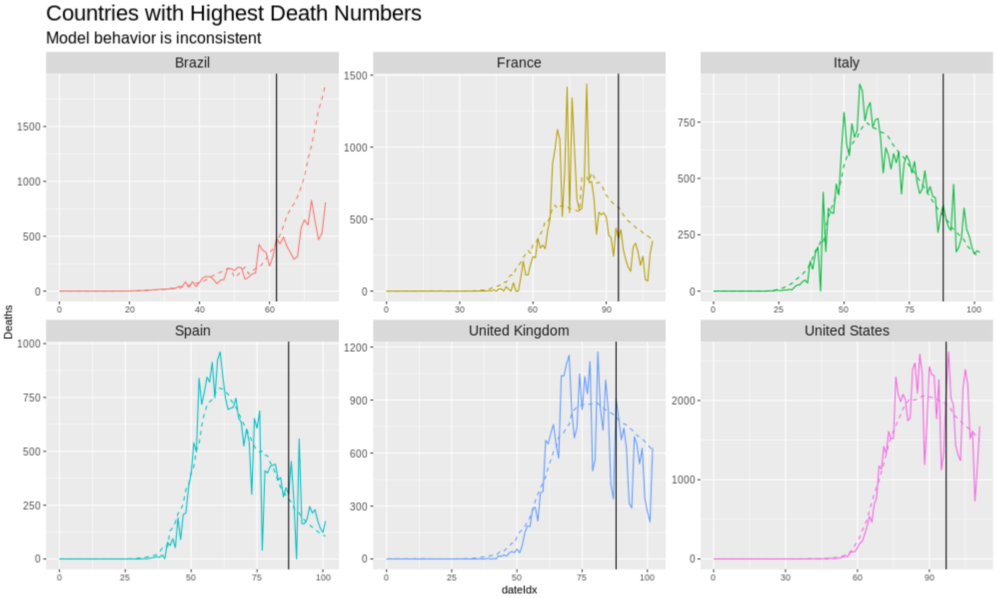
The countries the model performs best on are Russia, Italy, Peru, Saudi Arabia, the United States, and Colombia. In general, these countries hae a stable pattern in the death rates that is consistent between the training and evaluation data sets. For the countries with a larger number of deaths (Italy and the United States), the exact daily values have somewhat severe daily fluctuations; in these cases the model is consistent in keeping to the middle of the trend. The model obtains between a 17-23% prediction error for these countries.



The cases when the model performs worst are Sierra Leone, New Zealand, Senegal, Eswatini, Latvia, and Bahrain. Unsurprisingly, these countries have very few deaths in the valuation set, with 33 between all countries in the 15-day test period. Sierra Leone's total predictions error is 27 times larger the actual number of deaths. In this case, the over-prediction is due to a spike towards the end of the training data which, when played out over weeks, results in a sharp increase where none was actually present. I do not want to to overstate these results; between these 6 countries there were only 54 total deaths in the entire training set, from which is difficult to infer any signal. In addition, the remaining countries have so few deaths in the evaluation set that, even when the model is predicting fractions of a death per day, the will still greatly overestimate the number of deaths in the two week period. However this does highlight a limitation of the model in terms of usefulness when there isn't a stable pattern in the death rate established in the training data set.



The final case I will consider is the model's performance on the countries with the highest death total across all the available data: Italy, the United States, the United Kingdon, Spain, France, and Brazil (in order of performance). In general, the model correctly captures the general trajectory of the deaths; for Italy, Spain, and the United States the predictions are fairly accurate, obtaining 19%, 21%, and 32% errors respectively. On the other hand, the United Kingdom and France's rates decayed faster than the predicted values and Brazil did not increase nearly as dramatically as predicted (the model came close to over-predicting by 1,000 deaths per day). I'm unsure if these results arise from something similar to Sierra Leone's case where shifts near the train/evaluation threshold skew the predictions or if actual policy enactments substantially changed the death rate for Coronavirus patients in these countries.


Countries with Highest Death Numbers

Model behavior is inconsistent

Limitations and further work

There are two major limitations to this analysis: (1) the accuracy of the available data, (2) the limited feature set. A less inhibitive set-back but nonetheless something to be conscious of is (3) the process for estimating the death rates. For (1), the data on confirmed cases and recoveries are likely lower than the actual values as people in many countries are encouraged to stay at home and weather milder cases to avoid overstraining the hospital system. This results in an overestimation of the death rate, as deaths would still be accurately reported. For (2), this model only considers 'date since first confirmed case' and country as covariates. It is not unreasonable to think some quantification of 'healthcare availability' or 'policy/social measures taken' are missing from the model, and that these significantly influence the death rate. Furthermore, epidemic information broken down by age, socioeconomic status, and/or race would result in a model that more accurately reflects reality as the death rate is not homogeneous across different subpopulations. Finally for (3), from a usability standpoint needing to know the number of people who will be infected (and therefore at risk of dying from Coronavirus) at some later date requires another model run in tandem to estimate that quantity to feed into this model.

Given more time I would have liked to incorporating other variables into the model. Including data on healthcare availablilty is an obvious first choice, but from the census data it would be interesting to see if making use of the population-age distributions would yield better predictions. This is not an unreasonable hypothesis; a population that consists of more elderly (and likely high-risk) patients would either see more deaths or see the death rate increase more quickly (at least initially) due to this population. Along the same lines, it would have been interesting to see if countries that have a socialized medical system saw either less deaths, due to the availability of treatment, or had less of a disparity in outcomes due to socioeconomic status.

Exact Model Performance on Test Data

Table 2: Test Results for Best Performing Countries

Country	Total Deaths	Total Days	Total Error	Avg. Daily Error	Mean Abs. Deviance
Russia	1322	15	227.11735	15.141157	0.1717983
Italy	3934	15	759.50348	50.633565	0.1930614
Peru	1275	15	247.93836	16.529224	0.1944615
Saudi Arabia	120	15	23.42925	1.561950	0.1952437
United States	26137	15	5694.05799	379.603866	0.2178543
Colombia	240	15	55.85965	3.723976	0.2327485

Table 3: Test Results for Worst Performing Countries

Country	Total Deaths	Total Days	Total Error	Avg. Daily Error	Mean Abs. Deviance
Sierra Leone	15	15	412.284271	27.4856181	27.485618
New Zealand	2	15	17.244446	1.1496298	8.622223
Senegal	10	15	55.509918	3.7006612	5.550992
Eswatini	1	15	4.911005	0.3274003	4.911005
Latvia	5	15	18.339947	1.2226631	3.667989
Bahrain	1	15	3.583804	0.2389203	3.583804

Table 4: Test Results for Countries with Most Deaths

Country	Total Deaths	Total Days	Total Error	Avg. Daily Error	Mean Abs. Deviance
Italy	3934	15	759.5035	50.63357	0.1930614
United States	26137	15	5694.0580	379.60387	0.2178543
United Kingdom	8311	15	2656.0096	177.06731	0.3195776
Spain	3399	15	1268.4296	84.56198	0.3731773
France	3698	15	3106.1890	207.07927	0.8399646
Brazil	7858	15	8142.1141	542.80761	1.0361560