Does BCG vaccine have a protective effect against severe COVID-19?

Sounak Paul

December 9, 2020

Contents

1 Introduction

It has been suspected since the outbreak of SARS-CoV-2, that the bacillus Calmette–Guerin (BCG) vaccine might be providing some protection from the virus. The basis for this suspicion is primarily the observation that countries which had a national BCG vaccination policy seem to have lower number of cases and deaths per million people. The paper titled "BCG vaccine protection from severe coronavirus disease 2019 (COVID-19)" by Escobar et al., which appeared in PNAS , (https://doi.org/10.1073/pnas.2008410117) on October 12, 2020, claims that BCG policy and coverage indeed has a statistically significant effect on the number of

deaths per million in a country. In this project, I will comment on and criticize their methodology, and then conduct a much more sophisticated statistical analysis incorporating spatial and/or temporal structure in my models. In the end, I shall provide comments regarding reliability and other issues with my own analysis.

1.1 Brief discussion of original paper

Previous studies have suggested a negative association between national BCG vaccination policy and the prevalence and mortality of COVID-19. However, these studies are difficult to validate due to broad differences between countries such as socioeconomic status, demographic structure, rural vs. urban setting, time of arrival of the pandemic, number of diagnostic tests/criteria for testing, and national control strategies to limit the spread of COVID-19. This paper tries to address this issue by refining the epidemiological analysis to mitigate the effects of potentially confounding factors like stage of the COVID-19 epidemic, human development index, population density, age structure, etc, but they use very primitive methods to do so. That being said, no clinical trials were conducted in this study, and even the authors state that clinical trials are required to safely establish causality between BCG vaccination and protection from severe COVID-19.

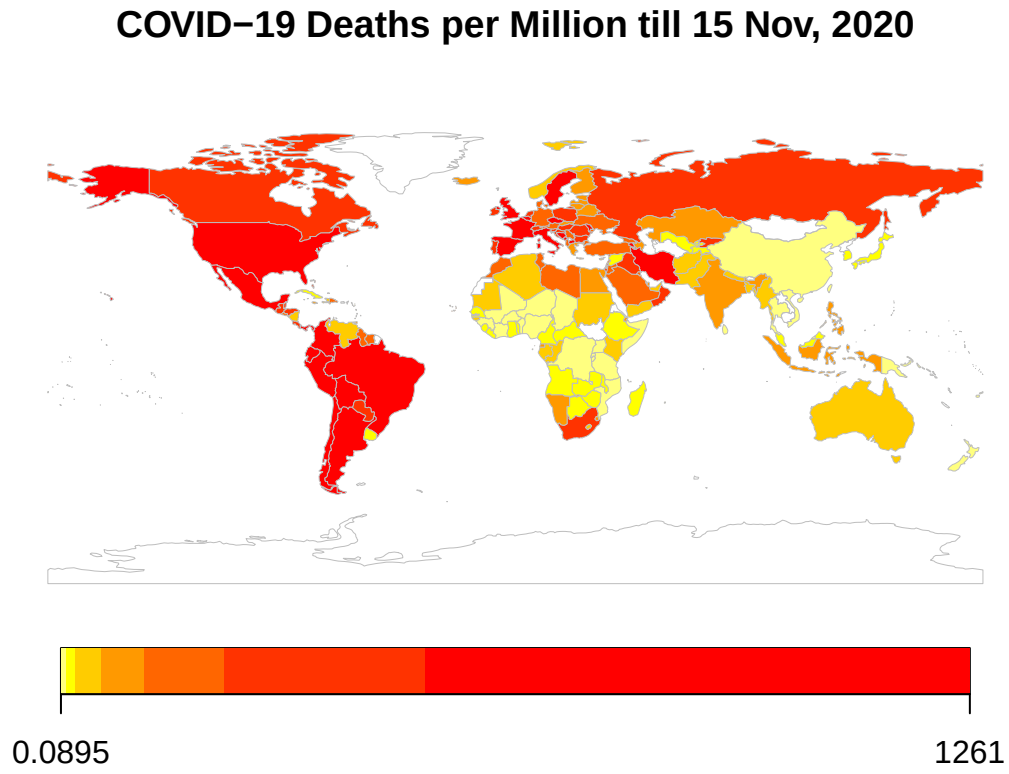
The main dataset associated with this paper, is a dataset the authors named "Coarse", and can be found here: (https://www.pnas.org/content/suppl/2020/07/07/2008410117.DCSupplemental). It contains several variables, the most important ones being Country (Identifier variable), region (Categorical with 7 levels), population in 2018, population density of 2018, urban percentage in 2018 (percentage of population living in cities), percentage of population aged 65 or above, and COVID-19 deaths per million in each of the first 4 weeks. It also contains BCG vaccination information in each country, through the columns "BCG Policy" (Categorical with 3 levels: current, interrupted and never), BCG vaccine start, end and year range, and BCG mean coverage. Using this, the authors mainly performed two analyses: a coarse analysis and a filtered/refined analysis (the tables and figures for which can be found in the appendix of the paper). The methods used for both these analyses were ANOVA, t-tests, and simple linear regression, which are crude at best. Even more questionable was the fact that a lot of linear regressions were done to test for strong associations between COVID-19 mortality and a large number of covariates, one by one. They found strong and consistent associations between HDI, percentage of population above 65, and urbanization, with COVID-19 mortality (Every 10 percent increase in the BCG index was associated with a 10.4 percent reduction in COVID-19 mortality). They also rejected the null hypothesis of no association between BCG vaccination and COVID-19 mortality. Issues of multiple testing that may arise due to so many tests, were completely ignored.

Their filtered analysis employed the exact same methods, the only difference being that here the authors tried to control for confounding variables by picking out 22 socially similar countries. Here, they added a new covariate, a BCG index, i.e. a proxy for the degree of universal BCG vaccination deployment in a country. While the control of confounding variables reduced the overall significance of the association between BCG and COVID-19 deaths, nevertheless, a significant association and effect was still detected for several comparisons controlling for social conditions and stage of the epidemic by country. They also do follow-up analyses of socially similar countries in Europe, again in an attempt to reduce confounders. There are a lot of issues of statistical interest that could have been addressed and explored, and much more sophisticated models using random effects and spatiotemporal kernels could have been deployed in the analysis, especially since the possibility of BCG vaccination having a protective effect against COVID-19 can have widespread ramifications.

1.2 My setup

In order to conduct a more sophisticated analysis by incorporating spatiotemporal correlations, I needed more data. Apart from the "Coarse" dataset available at the mentioned link, I used two more datasets: The first one contained country ISO3 codes along with their coordinates, i.e. latitude and longitude (https://gist.github.com/tadast/8827699), and the other was a time series dateset obtained from the Johns Hopkins Covid data repository on github (https://github.com/datasets/covid-19), containing the

cumulative number of deaths every day in each country, from 22nd January till 11th November (i.e. 299 days total). This is in stark contrast with the Coarse dataset, which used only the cumulative number of deaths after two weeks, and four weeks since the first death reported in each country. Using these three datsets, I created three other new datesets that I used in my anlysis (which can be viewed in my github repository, https://github.com/paulsounak96/stat349):



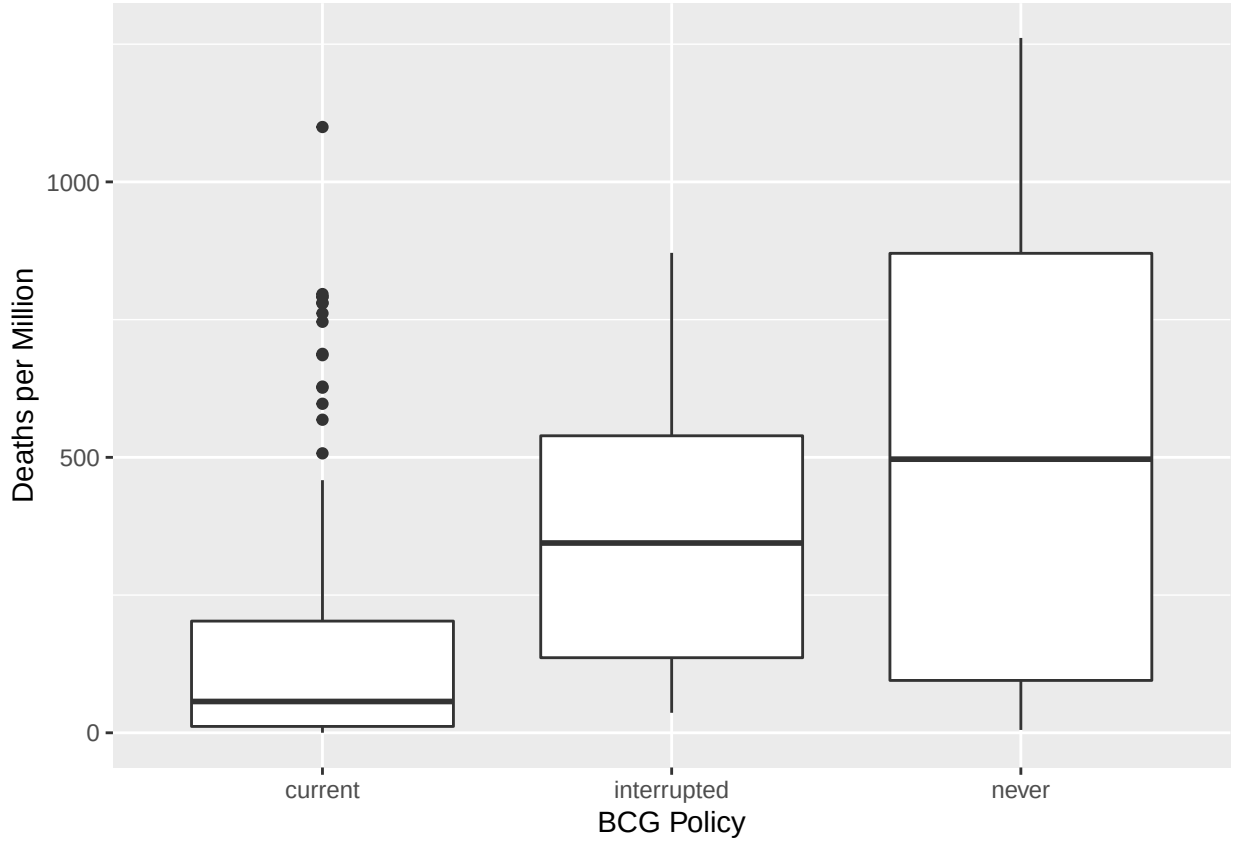
• **cumdailydeaths**: A dataset with 39550 observations of 19 columns, giving the cumulative daily deaths every day since the first confirmed death due to COVID-19 in each country (171 countries in total). This dataset was created by inner joining "Coarse" with the latitude and longitude dataset, and then with the JHU Covid set, after which some unnecessary covariates were removed. The observational units are (Country, Day), where Day has a temporal structure.

• **cumweeklydeaths**: A dataset with 5573 observations of 15 columns, giving the cumulative weekly deaths every day since the first confirmed death due to COVID-19 in each country (170 countries in total). This was made by aggregating dataset `cumdailydeaths` into blocks of 7 days, excluding the last few leftover days. The country Saint Lucia was excluded in this dataset since it had been just 6 days since its first covid death. The observational units are (Country, Week).

• **countryinfo**: A dataset with 171 observations of 14 columns, obtained by choosing the cumulative deaths in each each country on the final observed day. That gives us 171 observations corresponding to 171 countries.

Note that the "Coarse" dataset contained more than 300 entries, most of them well filled with missing values (and also, many of them were not actual countries, so it was tough to find latitude and longitude information, and daily/weekly covid death information about them). Also, the authors had considered each state of US as a different country arguing that many states had higher population than several countries. I do not believe that is a good thing to do, since as of November 15, we have several massive countries like India and Brazil, with huge populations as well as large number of deaths. So by the author's argument, we should break down these countries to their regions/states as well. But since finding the covid death records corresponding

to every region would be extremely time-consuming, I included only countries in my datasets. The only countries missing from my dataset are those which have either recorded zero deaths, or do not have reliable records of Covid deaths, or if it is unclear if BCG vaccine is administered in those countries. Even then, there were several missing values, especially in the column "range_age_BCG". Since we are mainly going to work with the dataset `cumweeklydeaths`, I shall provide a brief overview of its columns:



• Country: Categorical covariate, containing 171 levels, i.e. countries that have completed at least 1 week since their first recorded covid death.

• Week: Column recording number of weeks corresponding to every country, containing 171 levels, i.e. countries that have completed at least 1 week since their first recorded covid death. The highest number of weeks corresponding to a single country is 36 (will be regarded as both a factor as well as a continuous covariate).

• ISO3: Unique 3 digit identifier of every country (Factor variable again).

• population_million: Population of the country, in millions (Continuous covariate).

• HDI_2018: Human Development Index of corresponding country in 2018. Values are between 0 and 1.

• log_pop_density: Natural Logarithm of the population density of corresponding country. Log transformation was taken because this was a positive covariate, and the highest value was several orders of magnitude larger than the lowest. The original paper did not take this transformation.

• ages_65_up: Fraction of population above 65. In original dataset, this covariate was given as a percentage, but I considered it as a fraction, mainly for imputation of missing values via logistic regression, and also to ensure that that range of values were not too different from that of the response.

• BCG_policy: Factor variable with 3 levels: Current (for countries with ongoing universal BCG vaccination policy), Interrupted (for countries which previously had universal BCG vaccination policy, but was later abandoned).

• range_age_BCG: The range of ages of people BCG vaccine was administered on, in the country. The highest value attained by this variable is 92, for Romania.

• BCG_mean_coverage: The average of estimated fraction of infants that were administered the vaccine, across years. Few missing values present.

• urban_fraction: Fraction of population living in cities.

• lat: Latitude value of the corresponding country.

• long: Longitude value of the corresponding country.

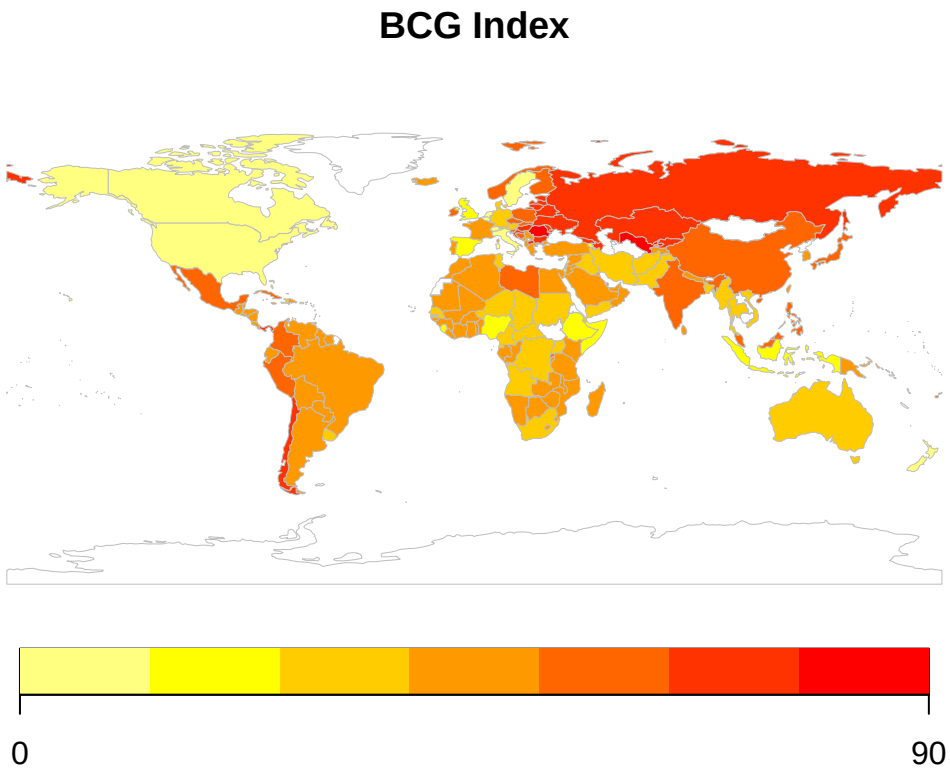
• Deaths: Cumulative number of deaths in a (Country, Week) pair.

• log_deaths_million: The logarithm of deaths per million population in a country. The log transformation was also taken in the original paper, and I have tried to justify it using the Box-Cox transformation (later on).

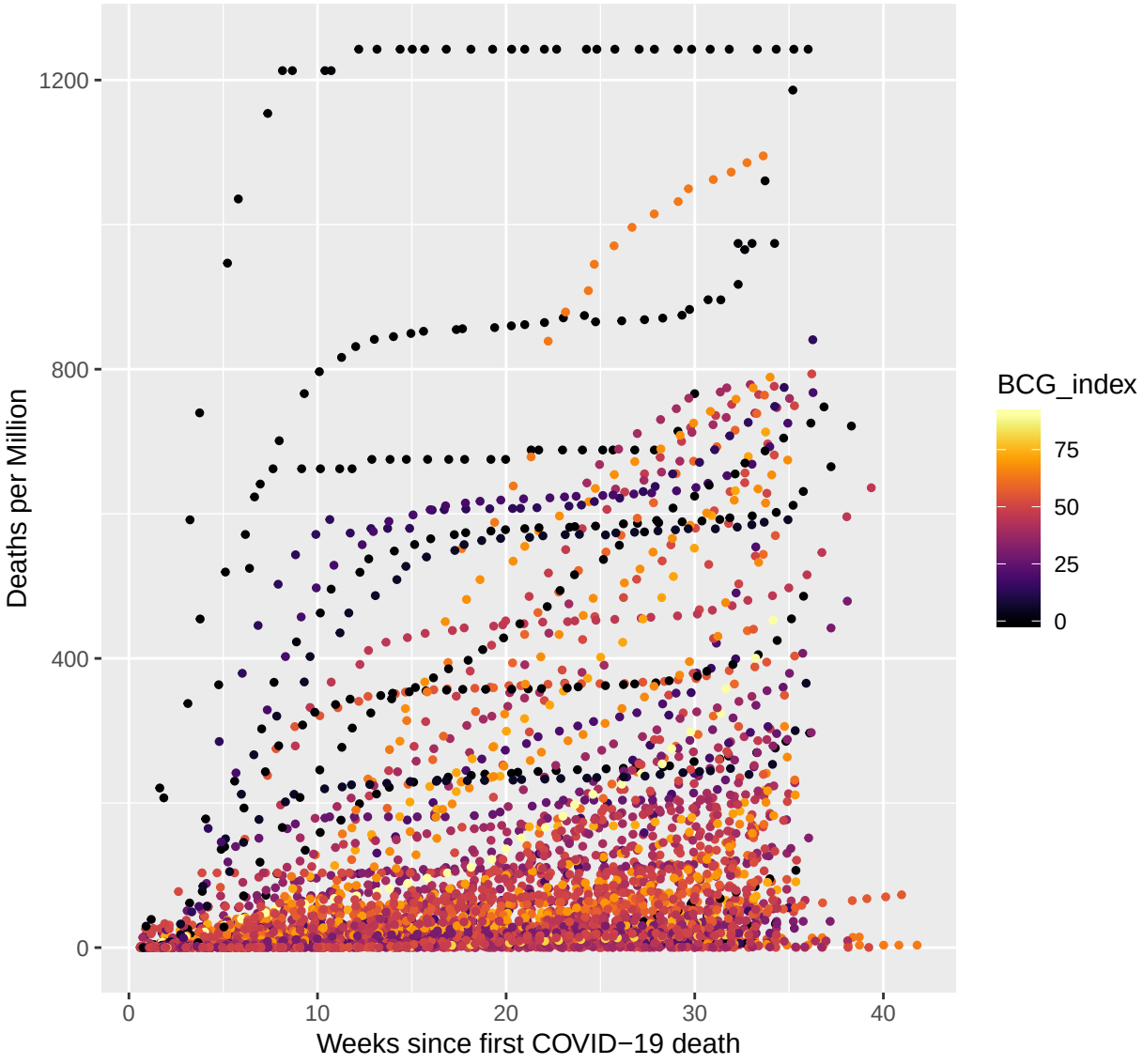
• BCG_index: A self defined measure of the BCG penetration in the country. The original paper also had its own definition of a BCG index, but I disagreed with it, and defined,

$$BCG\_index = range\_age\_BC \times BCG\_mean\_coverage\ .$$

BCG index is highest for Romania (equalling 89.96), since it has the highest value of $range\_age\_BC$, i.e. 92, and a very high BCG coverage of 97.8%.



The following diagram plots the number of deaths per million of every country across weeks since their respective first COVID-19 death, where the points are colored according to the BCG indices of their corresponding countries. Visually, countries with higher BCG index do seem to be recording lower deaths per million across weeks.



We now move on to our preliminary analysis, before fitting our models. Code is provided in the appendix.

2 Preliminary analysis

2.1 Imputation of missing values

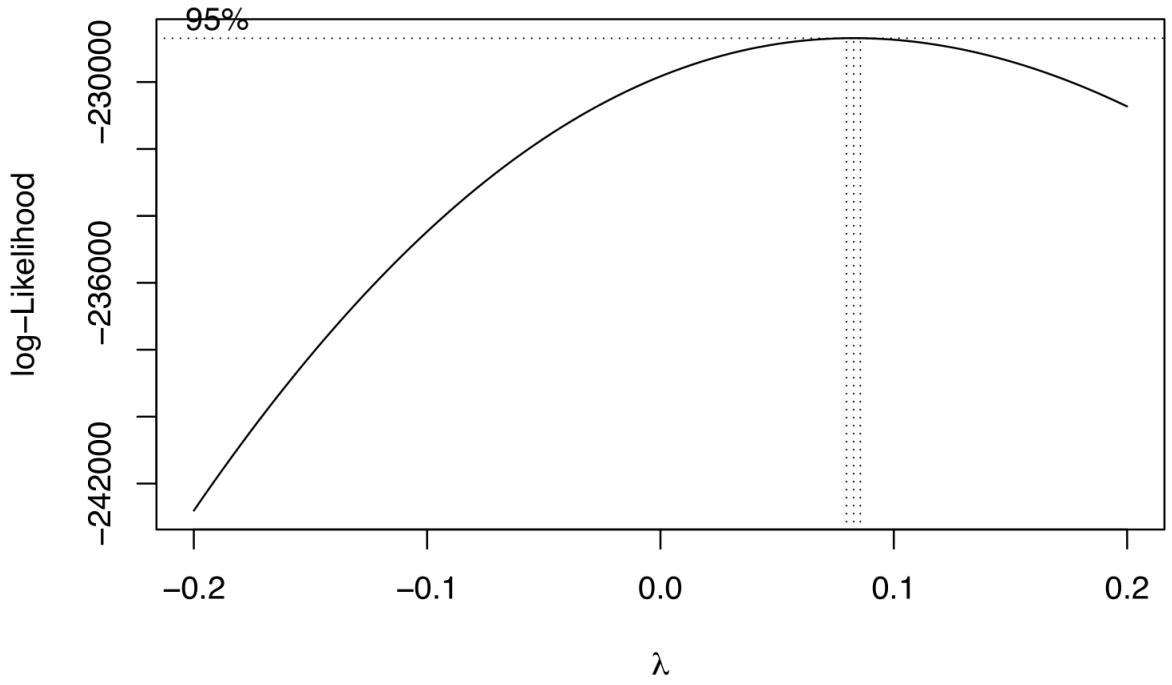
I searched the internet and tried to fill up missing values as much as possible, but some values were tough to find, especially regarding `ages_65_up` and `range_age_BCG`, and a few values of `BCG_mean_coverage`. For `range_age_BCG`, I set the values for countries with never had a BCG vaccination programme, to 0. For the rest, I imputed using the mean of the observed ranges of other countries, i.e. 47.8 years. I believe that this imputation does make sense, since most of the countries which have had a BCG vaccination programme at some point, have ranges well over a few decades. Since this will get multiplied with the mean BCG coverage over the years, to give the treatment, i.e. `BCG_index`, hence we do not need to worry too much in my opinion.

For `ages_65_up`, I imputed the missing values using the values predicted by logistic regression of `ages_65_up` on the covariates `HDI_2018`, `log_pop_density`, and `urban_fraction`. Finally, for `BCG_mean_coverage`, I

imputed the missing values again using the values predicted by logistic regression of `ages_65_up` on the covariates `ages_65_up`, `HDI_2018`, `log_pop_density`, and `urban_fraction`.

2.2 Transformation of response

Since the range of death per million is very huge, taking a log transformation was always an option. I used the `boxcox` function to find an optimal Box-Cox power transformation of the OLS regression model, where I regressed the number of deaths per million on `ages_65_up`, `HDI_2018`, `log_pop_density`, `urban_fraction`, and `BCG_index`. The plot of the log-likelihood against λ is given below.



While zero is not included in the 95% CI, the whole confidence interval was contained in (0, 0.1), which are extremely small values. Thus I believe that taking a log transformation would not be unreasonable. Also, I suspect that the effects of many of the covariates on deaths per million, could be multiplicative instead of additive. Hence I proceeded with the log transformation to obtain the response variable `log_deaths_million`.

2.3 t-test

Here, I conducted a very rudimentary test, i.e. the unpaired t-test. I took a subset of countries from the entire dataset for whom, at least 32 weeks had elapsed since the first death. The number 32 was chosen since it was the first quartile of all 171 countries. This created a new dataset `ttestdata`, which had 142 countries, their BCG policy, and the cumulative number of deaths on Day 228 since their first covid death.

Then, I ran a unpaired t-test with 2 samples, one with log deaths per million of countries that have never had a BCG vaccination policy, and the other with that of the rest. The results I obtained were as follows:

```
## [1] "t         2.0594"
```

```
## [1] "df        10.376"
```

```
## [1] "p-value   0.06544"
```

The 95% confidence interval for the difference in log deaths per million, of the 2 samples, is $(-0.094,\ 2.963)$. This shows that the difference is not statistically significant at the 0.05 level, though it is still quite close.

Note that this test does not take into account the BCG penetration in the countries, since neither the range of years of BCG administration, nor the mean coverage is used. Also, the t-test assumes independence of all the points in both samples. In this case however, the values of the response in neighbouring countries are spatially correlated. Thus, it is clear that spatiotemporal models are necessary to obtain more reliable results.

3 Fitted models

3.1 Without temporal component

In the previous section, since we had already created the dataset `ttestdata`, I think it would be worth fitting a model with the fixed effects plus only the spatial correlation.

- **Description of model:** The model I explored, (denoting the response `log_deaths_million` as Y) was

$$\begin{aligned} Y_c \ = \ \beta_0 \ + \ \beta_1 HDI_c \ + \ \beta_2 log\_pop\_density_c \ + \ \beta_3 ages\_65\_up_c \\ + \ \beta_4 BCG\_index_c \ + \ \beta_5 urban\_fraction_c \ + \ \epsilon_c \ + \ S_c \end{aligned} \tag{1}$$

where c denotes Country, $\epsilon_c \sim N(0, \sigma_0^2)$ is random noise independent for distinct countries, and

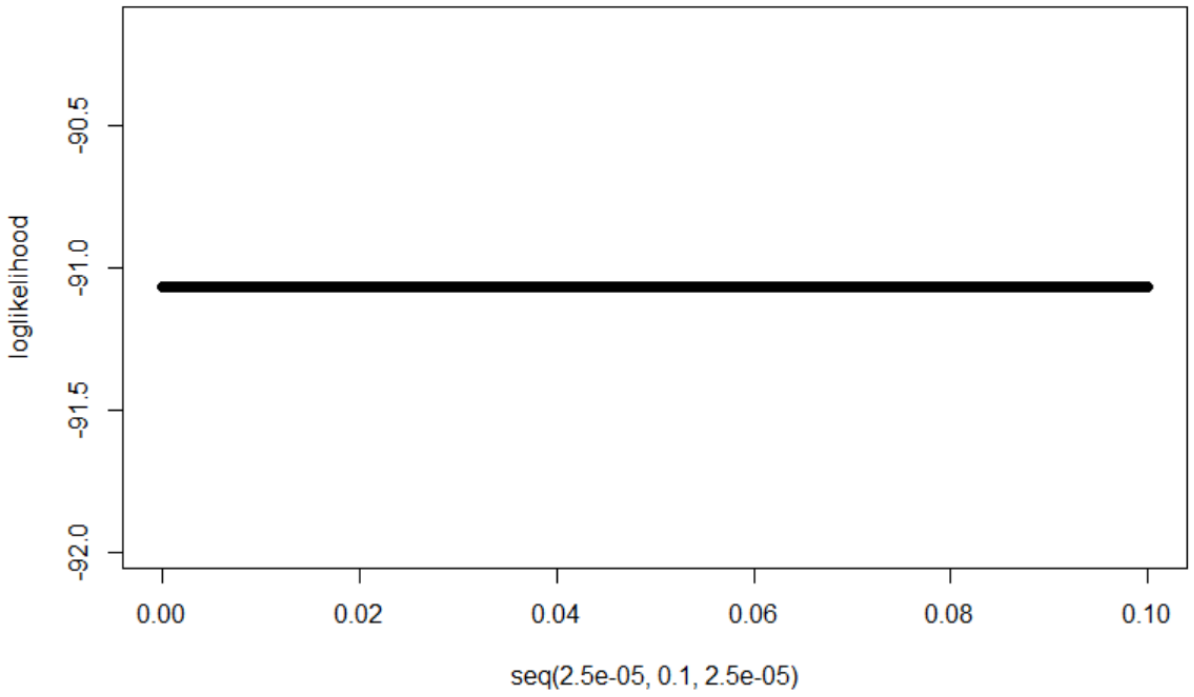
$$S \sim N(0, \sigma_1^2 \Sigma), \text{ where } \ \Sigma[i,j] = \exp\left(-\frac{\rho \|x_i - x_j\|}{20037508}\right)$$

is the exponential spatial covariance matrix. Here $\|.\|$ is the Haversine distance, i.e. the great-circle distance in metres between any 2 points on the earth, x_i is a longitude latitude tuple of the ith country, and ρ is a constant for scaling. Note that 20037508 metres is half the circumference of the earth (and the maximum possible distance between 2 points on it).

The optimum value ρ is found by plotting the log-likelihood of the above model (using REML) fitted without `BCG_index`, i.e.

$$l \ = \ -\frac{1}{2}\left[y'\hat{\Sigma}^{-1}Qy \ - \log\det(\hat{\Sigma}) \ - \log\det(X'\hat{\Sigma}^{-1}X) \ + \ \log\det(X'X)\right] \ ,$$

with the values of ρ going from 0.000025 to 0.1 with increments 0.000025. Note that in the above equation, Σ is a function of ρ, and Q is the orthogonal projection with kernel equalling the column space of X, the model matrix of the fixed effects. We see from the following plot, that the likelihood function stays constant across the entire interval under consideration.



While we could pick any non-zero constant ρ as an optimum value, we shall investigate the case where $\rho \to 0$ to see if we arrive at a simpler expression for the covariance.

- **Modification to model definition:** If we simply take $\rho \to 0$ in the definition of S, keeping σ_1^2 constant, then by the Taylor expansion of

$$\sigma_1^2 \exp\left(-\frac{\rho\|x_i - x_j\|}{20037508}\right) \;=\; \sigma_1^2 \;-\; \rho\sigma_1^2\frac{\|x_i - x_j\|}{20037508} \;+\; \rho^2\sigma_1^2\frac{\|x_i - x_j\|^2}{20037508^2} \;-\; \cdots,$$

we get that,

$$\sigma_1^2\Sigma[i,j] \;\to\; \sigma_1^2,$$

which is not very interesting to consider, since we expect the correlation of the responses corresponding to nearby countries to be greater than that of countries farther away. Things get much more interesting if we take $\rho \to 0$ keeping $\lambda = \rho\sigma_1^2$ constant, in which case we get

$$\sigma_1^2\Sigma[i,j] \;\to\; \sigma_1^2 \;-\; \lambda^2\frac{\|x_i - x_j\|}{20037508}.$$

Ignoring the σ^2 term, this enables us to simply consider

$$Z \sim N(0,\ \lambda^2\Sigma),\ \text{where}\ \ \Sigma[i,j] = -\frac{\|x_i - x_j\|}{20037508} \tag{2}$$

in model (1), keeping everything else unchanged. Notice that this is simply the expresiion of the covariance function for Brownian motion on contrasts. We have still kept the denominator of 20037508 to ensure that the elements of Σ do not become too large, otherwise we might get extremely small estimates of λ.

- **Parameter estimates:** The parameter estimates along with their standard errors are

Parameter	Corresponding to	Estimate	Std.Error
β_0	Intercept	3.768	-
β_1	HDI	1.296	1.328
β_2	log_pop_density	-0.096	0.079
β_3	ages_65_up	-7.924	3.015
β_4	BCG_index	-0.010	0.005
β_5	urban_fraction	0.985	0.662
σ_0^2	ϵ	0.748	0.154
λ^2	Σ	9.383	3.191

Recall that BCG index was continuous variable between 0 to 90, defined as the product of the range of ages BCG vaccine was administered on, and the average of estimated fraction of infants that it was administered on. The mean of the logarithm of deaths per million decreases by 0.010 with a standard error of 0.005, for every increase in BCG index by 1. We can see that the standard error of BCG index is half of its parameter estimate, which suggests that the effect of BCG index is borderline significant.

• **Significance of BCG index:** Now we fit the model without and with the treatment `BCG_index`, taking the kernel to be the the subspace spanned by the fixed effects of the former model. The likelihood ratio chi square statistic of this model with its reduced one, turns out to be 3.05017 on 1 degree of freedom, which gives the corresponding p-value as 0.081. Thus, from this analysis we get that the effect of `BCG_index` is not statistically significant.

3.2 Spatiotemporal models

Now, we shall fit our first model on the `cumweeklydeaths` dataset which would incorporate spatiotemporal kernels.

3.2.1 Brownian motion on simple contrasts for both space and time

• **Description of model:** Consider the model,

$$\begin{aligned} Y_{ct} \;=\; \beta_0 \;+\; \beta_1 t \;+\; \beta_2 HDI_c \;+\; \beta_3 log\_pop\_density_c \;+\; \beta_4 BCG\_index_c \\ +\; \beta_5 ages\_65\_up_c \;+\; \beta_6 urban\_fraction_c \;+\; \epsilon_{ct} \;+\; S_c \;+\; T_t, \end{aligned} \tag{3}$$

where c and t refer to country and time (in weeks), $\epsilon_c \sim N(0,\ \sigma_0^2)$ is random noise irrespective of time and country effect,

$$S \sim N(0,\ \sigma_1^2 \Sigma_{space}),\ \text{where}\ \ \Sigma_{space}[i(c_1,t_1),\ j(c_2,t_2)] = -\frac{\|x_{c_1} - x_{c_2}\|}{20037508}\ ,$$

$$T \sim N(0,\ \sigma_2^2 \Sigma_{time}),\ \text{where}\ \ \Sigma_{time}[i(c_1,t_1),\ j(c_2,t_2)] = -|t_1 - t_2|\ .$$

Here, σ_1 can be thought of as a volatility coeffiecient that is common across all weeks, and σ_2 across all countries. $\|.\|$ is the Haversine distance as before, and i and j are row and column indices of Σ corresponding to country c_1 at time t_1, and country c_2 at time t_2 respectively. Notice that both the spatial and temporal covariance functions are of Brownian motion on simple contrasts. One advantage of this choice is that we do not need to introduce scale factors, unlike in exponential covariance functions.

• **Parameter estimates:** The parameter estimates along with their standard errors are:

Parameter	Corresponding to	Estimate	Std.Error
β_0	Intercept	-3.552	-
β_1	Time (in Weeks)	0.286	0.081
β_2	HDI	2.355	1.318
β_3	log_pop_density	0.046	0.127
β_4	BCG_index	-0.026	0.006
β_5	ages_65_up	2.884	3.177
β_6	urban_fraction	0.972	0.623
σ_0^2	ϵ	0.623	0.024
σ_1^2	Σ_{space}	40.024	6.111
σ_2^2	Σ_{time}	0.045	0.020

The mean of the logarithm of deaths per million decreases by 0.026 with a standard error of 0.006, for every increase in BCG index by 1.

• **Significance of BCG index:** As before, we fit the model without and with the treatment `BCG_index`, taking the kernel to be the the subspace spanned by the fixed effects of the former model. The likelihood ratio chi square statistic of this model with its reduced one, turns out to be 25.96 on 1 degree of freedom, which gives us a statistically very significant p-value of 3.486×10^{-7}. This is the first model where we get that the effect of `BCG_index` is statistically significant, unlike the previous model which did not incorporate the temporal structure.

3.2.2 Inclusion of interaction of spatial and temporal effects

Our final model is the most general one, and the only difference from the previous model is that this model has one more variance component, i.e. the interaction of the spatial and temporal effect. Hence I will not give detailed explanation of the meaning of all the terms since they are the same as before.

• **Description of model:** Consider the model,

$$\begin{aligned} Y_{ct} \;=\; \beta_0 \;+\; \beta_1 t \;+\; \beta_2 HDI_c \;+\; \beta_3 log\_pop\_density_c \;+\; \beta_4 BCG\_index_c \\ +\; \beta_5 ages\_65\_up_c \;+\; \beta_6 urban\_fraction_c \;+\; \epsilon_{ct} \;+\; S_c \;+\; T_t + \; Z_{ct}, \end{aligned} \tag{4}$$

where $\epsilon_c \sim N(0, \sigma_0^2)$ is random noise irrespective of time and country effect,

$$\begin{aligned} S \sim N(0,\; \sigma_1^2 \Sigma_{space}),\; \text{where}\;\; \Sigma_{space}[i(c_1,t_1),\; j(c_2,t_2)] = -\frac{\|x_{c_1} - x_{c_2}\|}{20037508}\;, \\ T \sim N(0,\; \sigma_2^2 \Sigma_{time}),\; \text{where}\;\; \Sigma_{time}[i(c_1,t_1),\; j(c_2,t_2)] = -|t_1 - t_2|\;, \\ Z \sim N(0,\; \sigma_3^2 \Sigma_{st}),\; where\;\; \Sigma_{st}[i(c_1,t_1),\; j(c_2,t_2)] = \exp\left(-\frac{\|x_{c_1} - x_{c_2}\|}{20037508} - |t_1 - t_2|\right)\;. \end{aligned} \tag{5}$$

Note that Σ_{st} is simply the Hadamard product of $\exp(\Sigma_{space})$ and $\exp(\Sigma_{time})$, where the exponents are taken entry-wise. I believe this model intuitively makes the most sense, since in addition to σ_1 and σ_2 from the previous model, here σ_3 is a volatility coefficient for the interaction of spatial and temporal effects. The exponent has been taken to ensure that the Hadamard product indeed returns a positive semidefinite function.

It is worth mentioning that it would have made more sense to define

$$\Sigma_{st}[i(c_1,t_1),\; j(c_2,t_2)] = \exp\left(-\rho_1 \frac{\|x_{c_1} - x_{c_2}\|}{20037508} - \rho_2 |t_1 - t_2|\right)\;, \tag{6}$$

and find the optimum value of ρ which maximizes the REML log-likelihood (same procedure as in Subsection 3.1). However, this model takes a really long time to run even once, hence it would not have been feasible to run the corresponding code for different values of ρ_1 and ρ_2. Hence, I stuck to the definition in (5).

• **Modification to model definition:** When we try to fit the above model, we end up getting a negative parameter estimate of σ_2^2 (corresponding to Σ_{time}), i.e. -0.406 with a standard error of 0.297. Thus, we shall remove the random variable T from the model definition in equation (4) and fit the following model:

$$Y_{ct} = \beta_0 + \beta_1 t + \beta_2 HDI_c + \beta_3 log\_pop\_density_c + \beta_4 BCG\_index_c + \beta_5 ages\_65\_up_c + \beta_6 urban\_fraction_c + \epsilon_{ct} + S_c + Z_{ct}, \quad (6)$$

where the terms are as defined before.

• **Parameter estimates:** The parameter estimates for the model in (6), along with their standard errors are:

Parameter	Corresponding to	Estimate	Std.Error
β_0	Intercept	-2.466	-
β_1	Time (in Weeks)	0.244	0.127
β_2	HDI	2.347	1.319
β_3	log_pop_density	0.060	0.130
β_4	BCG_index	-0.025	0.006
β_5	ages_65_up	2.669	3.267
β_6	urban_fraction	0.980	0.626
σ_0^2	ϵ	0.204	0.015
σ_1^2	Σ_{space}	41.702	6.312
σ_3^2	Σ_{st}	4.154	0.353

The mean of the logarithm of deaths per million decreases by 0.025 with a standard error of 0.006, for every increase in BCG index by 1.

• **Significance of BCG index:** Here also, we fit the model without and with the treatment `BCG_index`, taking the kernel to be the the subspace spanned by the fixed effects of the former model. The likelihood ratio chi square statistic of this model with its reduced one, turns out to be 42.26 on 1 degree of freedom, which gives us a statistically very significant p-value of 7.992×10^{-11}. In fact, according to this model, the effect of `BCG_index` is even more significant the the previous ones. In fact, the log likelihood of this model is 152.49 more than the model in the previous subsection, i.e. without the Σ_{st} interaction term. Hence we see that this model provides the best fit among all our models.

4 Final thoughts

We conclude from our spatiotemporal models, that the effect of BCG index on the logarithm of deaths per million is indeed significant, which agrees with the result of the original paper, albeit this time with a much deeper statistical analysis. That being said, there are some issues to keep in mind.

1. **Limited computational resources:** As mentioned before, we could not pursue the model with Σ_{st} defined as in equation (6), because of the long computational runtime needed to find the best hyperparameters ρ_1 and ρ_2 for Σ_{st}. Working with the dataset 'cumdailydeaths' (containing daily COVID-19 death information) was also next to impossible because of the massive size of the dataset, leading to a long runtime. Thus, an even more thorough analysis can be conducted using the same models used in this report, but using 'cumdailydeaths' instead of 'cumweeklydeaths' if superior computational resources are available. While I feel it is unlikely that the main results of my report would change, it would make the results more reliable.

2. **Calculation of mean vaccine coverage:** In my opinion, the weighted mean of BCG vaccine coverage should have been considered instead of BCG mean coverage, since the population of some countries have risen dramatically since the introduction of universal BCG vaccination programme. For example, India's population has almost quadrupled since 1948, the year BCG vaccination programme began in India. Thus it would make more sense to consider the average of BCG vaccine coverage weighted by the population of the country in that year. This could have been addressed with

3. **Ecological fallacy and aggregation bias:** Clinical trials are what should generally be used to make inferences about individuals, but in this case, we are using countries as our observational units. Hence this is an observational study, and the results could be a case of ecological fallacy, since observed relationships for groups might not necessarily hold for individuals. For example, Simpson's paradox is a classic ecological fallacy, where during comparison of two populations broken up in groups of (highly) unequal sizes, the mean of some variable could be higher in every group for the first population, but still be lower for the total population. Aggregate data often easier to obtain than data on individuals, hence ecological inferences are often made. However, confounding and aggregation bias often make the results unreliable.

The third point is in particular, the most alarming. Even if clinical trials fail to establish causality of increased COVID-19 protection due to BCG vaccine, I would argue that it still does not make my analysis completely redundant. My methods required no monetary resources, modest computational resources, and very short time of execution. They do indicate that the question of BCG vaccine offering improved resistance against COVID-19 is indeed worth further investigation by clinical trials, which are substantially more time consuming as well as costly. Hence, my analysis can be thought of as an important intermediate step.

5 References

- L. E. Escobar, A. Molina-Cruz, C. Barillas-Mury, BCG vaccine protection from severe coronavirus disease 2019 (COVID-19). *Proc. Natl. Acad. Sci. U.S.A.* **117**, 17720–17726 (2020).
- Johns Hopkins University Center for Systems Science and Engineering (CSSE), COVID-19 dataset, *GitHub* https://github.com/datasets/covid-19 (2020).
- T. Tamošauskas, Countries with their (ISO 3166-1) Alpha-2 code, Alpha-3 code, UN M49, average latitude and longitude coordinates. *GitHubGist* https://gist.github.com/tadast/8827699 (2014).
- S. Paul, *GitHub*, https://github.com/paulsounak96/stat349). (2020)

6 Appendix

```
load("cumdailydeaths.Rda")
load("cumweeklydeaths.Rda")

countryinfo = cumdailydeaths %>%
  group_by(Country, population_million, HDI_2018, log_pop_density, ages_65_up,
           BCG_policy, range_age_BCG, BCG_mean_coverage, urban_fraction, lat, long) %>%
  summarise(Days_since_first_death = n(),
            max_log_deaths_million = max(log_deaths_million))
countryinfo = as.data.frame(countryinfo)
head(countryinfo[, 1:5])

###########Imputations of BCG Age range###########
```

```r
cumdailydeaths$range_age_BCG[cumdailydeaths$BCG_policy == "never"] = 0
cumdailydeaths$range_age_BCG[cumdailydeaths$BCG_policy != "never"] =
  mean(countryinfo$range_age_BCG[cumdailydeaths$BCG_policy != "never"], na.rm = TRUE)

cumweeklydeaths$range_age_BCG[cumweeklydeaths$BCG_policy == "never"] = 0
cumweeklydeaths$range_age_BCG[cumweeklydeaths$BCG_policy != "never"] =
  mean(countryinfo$range_age_BCG[cumweeklydeaths$BCG_policy != "never"], na.rm = TRUE)

countryinfo$range_age_BCG[countryinfo$BCG_policy == "never"] = 0
countryinfo$range_age_BCG[countryinfo$BCG_policy != "never"] =
  mean(countryinfo$range_age_BCG[cumweeklydeaths$BCG_policy != "never"], na.rm = TRUE)

###########Imputations of ages_65_up###########

mod_ages65up = glm(ages_65_up ~ HDI_2018 + log_pop_density + urban_fraction,
                   data = countryinfo, family = binomial())

for (country in countryinfo$Country[is.na(countryinfo$ages_65_up)]) {
  for (i in 1:171) {
    if (countryinfo$Country[i] == country) {
      countryinfo$ages_65_up[i] = predict(mod_ages65up,
                data.frame(HDI_2018 = countryinfo$HDI_2018[i],
                           log_pop_density = countryinfo$log_pop_density[i],
                           urban_fraction = countryinfo$urban_fraction[i]),
                type = "response")
    }
  }
  for (j in 1:39550) {
    if (cumdailydeaths$Country[j] == country) {
      cumdailydeaths$ages_65_up[j] = predict(mod_ages65up,
                data.frame(HDI_2018 = cumdailydeaths$HDI_2018[j],
                           log_pop_density = cumdailydeaths$log_pop_density[j],
                           urban_fraction = cumdailydeaths$urban_fraction[j]),
                type = "response")
    }
  }
  for (k in 1:5573) {
    if (cumweeklydeaths$Country[k] == country) {
      cumweeklydeaths$ages_65_up[k] = predict(mod_ages65up,
                data.frame(HDI_2018 = cumweeklydeaths$HDI_2018[k],
                           log_pop_density = cumweeklydeaths$log_pop_density[k],
                           urban_fraction = cumweeklydeaths$urban_fraction[k]),
                type = "response")
    }
  }
}

###########Imputations of BCG mean coverage###########

mod_bcgcover = glm(BCG_mean_coverage ~ ages_65_up + HDI_2018 + log_pop_density +
                     urban_fraction, data = countryinfo, family = binomial())

for (country in countryinfo$Country[is.na(countryinfo$BCG_mean_coverage)]) {
```

```
  for (i in 1:171) {
    if (countryinfo$Country[i] == country) {
      countryinfo$BCG_mean_coverage[i] = predict(mod_bcgcover,
                data.frame(ages_65_up = countryinfo$ages_65_up[i],
                           HDI_2018 = countryinfo$HDI_2018[i],
                           log_pop_density = countryinfo$log_pop_density[i],
                           urban_fraction = countryinfo$urban_fraction[i]),
                type = "response")
    }
  }
  for (j in 1:39550) {
    if (cumdailydeaths$Country[j] == country) {
      cumdailydeaths$BCG_mean_coverage[j] = predict(mod_bcgcover,
                data.frame(ages_65_up = cumdailydeaths$ages_65_up[k],
                           HDI_2018 = cumdailydeaths$HDI_2018[j],
                           log_pop_density = cumdailydeaths$log_pop_density[j],
                           urban_fraction = cumdailydeaths$urban_fraction[j]),
                type = "response")
    }
  }
  for (k in 1:5573) {
    if (cumweeklydeaths$Country[k] == country) {
      cumweeklydeaths$BCG_mean_coverage[k] = predict(mod_bcgcover,
                data.frame(ages_65_up = cumweeklydeaths$ages_65_up[k],
                           HDI_2018 = cumweeklydeaths$HDI_2018[k],
                           log_pop_density = cumweeklydeaths$log_pop_density[k],
                           urban_fraction = cumweeklydeaths$urban_fraction[k]),
                type = "response")
    }
  }
}

countryinfo$BCG_index = countryinfo$BCG_mean_coverage * countryinfo$range_age_BCG
cumdailydeaths$BCG_index =
  cumdailydeaths$BCG_mean_coverage * cumdailydeaths$range_age_BCG
cumweeklydeaths$BCG_index =
  cumweeklydeaths$BCG_mean_coverage * cumweeklydeaths$range_age_BCG


##############Box Cox Transformation#########################


boxcox_mod = lm(Deaths/population_million ~ as.numeric(Day) + HDI_2018 +
                  log_pop_density + ages_65_up + BCG_index + urban_fraction,
                data = cumdailydeaths)
boxcox(boxcox_mod, seq(-0.2, 0.2, by = 0.01))


###################t test#################################


ttestdata = cumweeklydeaths %>%
  group_by(Country, ISO3, population_million, HDI_2018, log_pop_density,
```

```
          ages_65_up, BCG_policy, BCG_mean_coverage, range_age_BCG, BCG_index,
          urban_fraction, lat, long) %>%
  summarise(deaths = Deaths[32], response = log_deaths_million[32])
ttestdata = as.data.frame(ttestdata[!is.na(ttestdata$response), ])

paste0("Current = ", sum(ttestdata$BCG_policy == "current"),
      ", Interrupted = ", sum(ttestdata$BCG_policy == "interrupted"),
      ", Never = ", sum(ttestdata$BCG_policy == "never"))
t.test(ttestdata$response[ttestdata$BCG_policy == "current"],
       ttestdata$response[ttestdata$BCG_policy != "current"])
t.test(ttestdata$response[ttestdata$BCG_policy == "never"],
       ttestdata$response[ttestdata$BCG_policy != "never"])


##########################No Temporal Component#####################


save(countryinfo, file = "ctryin.Rda")
save(cumdailydeaths, file = "cddeaths.Rda")
save(cumweeklydeaths, file = "cwdeaths.Rda")

spatial_cov = matrix(1, 142, 142)
for (i in 1:142) {
  for (j in 1:142) {
    spatial_cov[i,j]=exp(-distHaversine(c(ttestdata$long[i],ttestdata$lat[i]),
                                        c(ttestdata$long[j],ttestdata$lat[j]))/20037508)
  }
}

x0 = model.matrix(~ ttestdata$HDI_2018 + ttestdata$log_pop_density +
                    ttestdata$ages_65_up + ttestdata$urban_fraction)

loglikelihood = c()

for (rho in seq(0.000025, 0.1, 0.000025)) {
  m = spatial_cov^rho
  mod_notemp0 = regress(response ~ HDI_2018 + log_pop_density + ages_65_up +
                          urban_fraction, ~m,
                        start = mod_notemp0$sigma, data = ttestdata)
  loglikelihood = c(loglikelihood, mod_notemp0$llik)
}
plot(seq(0.000025, 0.1, 0.000025), loglikelihood)

############################Spatial Only################################

spatial_cov1 = matrix(1, 142, 142)
for (i in 1:142) {
  for (j in 1:142) {
    spatial_cov1[i,j]=-distHaversine(c(ttestdata$long[i],ttestdata$lat[i]),
                                     c(ttestdata$long[j],ttestdata$lat[j]))/20037508
  }
}
```

```
x0 = model.matrix(~ ttestdata$HDI_2018 + ttestdata$log_pop_density +
                    ttestdata$ages_65_up + ttestdata$urban_fraction)

mod_ttest0 = regress(response ~ HDI_2018 + log_pop_density + ages_65_up +
                         urban_fraction, ~spatial_cov1,
                       start = mod_notemp0$sigma, data = ttestdata)

mod_ttest1a = regress(response ~ HDI_2018 + log_pop_density + ages_65_up +
                        urban_fraction + BCG_index, ~spatial_cov1, kernel = x0,
                      start = mod_notemp0$sigma, data = ttestdata)

2*(mod_ttest1a$llik - mod_ttest0$llik)
pchisq(2*(mod_ttest1a$llik - mod_ttest0$llik), 1, lower.tail = FALSE)

mod_ttest1b = regress(response ~ HDI_2018 + log_pop_density + ages_65_up +
                        BCG_index + urban_fraction, ~spatial_cov1,
                      start = mod_ttest0$sigma, data = ttestdata)
summary(mod_ttest1b)

######################Spatiotemporal Models Setup##############################

cumweeklydeaths=inner_join(data.frame(id = 1:170, Country= countryinfo$Country),
                           cumweeklydeaths, by = "Country")

sm = matrix(1, 170, 170)
exp_sm = matrix(1, 170, 170)
spatial_cov = matrix(1, 5573, 5573)
exp_spatial_cov = matrix(1, 5573, 5573)

for (i in 1:170) {
  for (j in 1:170) {
    sm[i,j] = -distHaversine(c(countryinfo$long[i],countryinfo$lat[i]),
                                  c(countryinfo$long[j],countryinfo$lat[j]))/20037508
  }
}
for (i in 1:170) {
  for (j in 1:170) {
    exp_sm[i,j] = exp(-distHaversine(c(countryinfo$long[i],countryinfo$lat[i]),
                             c(countryinfo$long[j],countryinfo$lat[j]))/20037508)
  }
}
for (i in 1:5573) {
  for (j in 1:5573) {
    spatial_cov[i,j] = sm[cumweeklydeaths$id[i], cumweeklydeaths$id[j]]
  }
}
for (i in 1:5573) {
  for (j in 1:5573) {
    exp_spatial_cov[i,j] = exp_sm[cumweeklydeaths$id[i], cumweeklydeaths$id[j]]
  }
}

temporal_cov = -abs(outer(cumweeklydeaths$Week, cumweeklydeaths$Week, "-"))
```

```
st_cov = exp_spatial_cov*exp(temporal_cov)

x0 = model.matrix(~ cumweeklydeaths$Week + cumweeklydeaths$HDI_2018 +
                    cumweeklydeaths$log_pop_density +
                    cumweeklydeaths$ages_65_up + cumweeklydeaths$urban_fraction)

######################Brownian motion on contrasts#################################

mod_notemp2 = regress(log_deaths_million ~ Week + HDI_2018 + log_pop_density +
                        ages_65_up + urban_fraction, ~ spatial_cov + temporal_cov,
                      tol = 0.001, data = cumweeklydeaths)

mod_notemp3a = regress(log_deaths_million ~ Week + HDI_2018 + log_pop_density +
                         BCG_index + ages_65_up + urban_fraction,
                       ~ spatial_cov + temporal_cov,
                       kernel = x0, start = mod_notemp2$sigma,
                       tol = 0.001, data = cumweeklydeaths)

2*(mod_notemp3a$llik - mod_notemp2$llik)
pchisq(2*(mod_notemp3a$llik - mod_notemp2$llik), 1, lower.tail = FALSE)

mod_notemp3b = regress(log_deaths_million ~ Week + HDI_2018 + log_pop_density +
                         BCG_index + ages_65_up + urban_fraction,
                       ~ spatial_cov + temporal_cov, start = mod_notemp2$sigma,
                       tol = 0.001, data = cumweeklydeaths)
summary(mod_notemp3b)

#########################Including Interactions#####################################

mod_notemp4 = regress(log_deaths_million ~ Week + HDI_2018 + log_pop_density +
                        ages_65_up + urban_fraction, ~ spatial_cov + temporal_cov + st_cov,
                      tol = 0.001, data = cumweeklydeaths)

mod_notemp5a = regress(log_deaths_million ~ Week + HDI_2018 + log_pop_density +
                         BCG_index + ages_65_up + urban_fraction,
                       ~ spatial_cov + temporal_cov + st_cov,
                       kernel = x0, start = mod_notemp4$sigma,
                       tol = 0.001, data = cumweeklydeaths)

2*(mod_notemp5a$llik - mod_notemp4$llik)
pchisq(2*(mod_notemp5a$llik - mod_notemp4$llik), 1, lower.tail = FALSE)

mod_notemp5b = regress(log_deaths_million ~ Week + HDI_2018 + log_pop_density +
                         BCG_index + ages_65_up + urban_fraction,
                       ~ spatial_cov + temporal_cov + st_cov,
                       start = mod_notemp4$sigma,
                       tol = 0.001, data = cumweeklydeaths)
summary(mod_notemp5b)

#########################Final model (without temporal)##############################

mod_notemp6 = regress(log_deaths_million ~ Week + HDI_2018 + log_pop_density +
                      ages_65_up + urban_fraction, ~ spatial_cov + st_cov,
```

```
                      tol = 0.001, data = cumweeklydeaths)

mod_notemp7a = regress(log_deaths_million ~ Week + HDI_2018 + log_pop_density +
                    BCG_index + ages_65_up + urban_fraction, ~ spatial_cov + st_cov,
                    kernel = x0, start = mod_notemp6$sigma, tol = 0.001, data = cumweeklydeaths)

2*(mod_notemp7a$llik - mod_notemp6$llik)
pchisq(2*(mod_notemp7a$llik - mod_notemp6$llik), 1, lower.tail = FALSE)

mod_notemp7b = regress(log_deaths_million ~ Week + HDI_2018 + log_pop_density +
                    BCG_index + ages_65_up + urban_fraction, ~ spatial_cov + st_cov,
                    start = mod_notemp6$sigma, tol = 0.001, data = cumweeklydeaths)
summary(mod_notemp7b)
```