BCG vaccine protection from severe coronavirus disease 2019 (COVID-19)

Sounak Paul

University of Chicago

paulsounak96@uchicago.edu

November 30, 2020

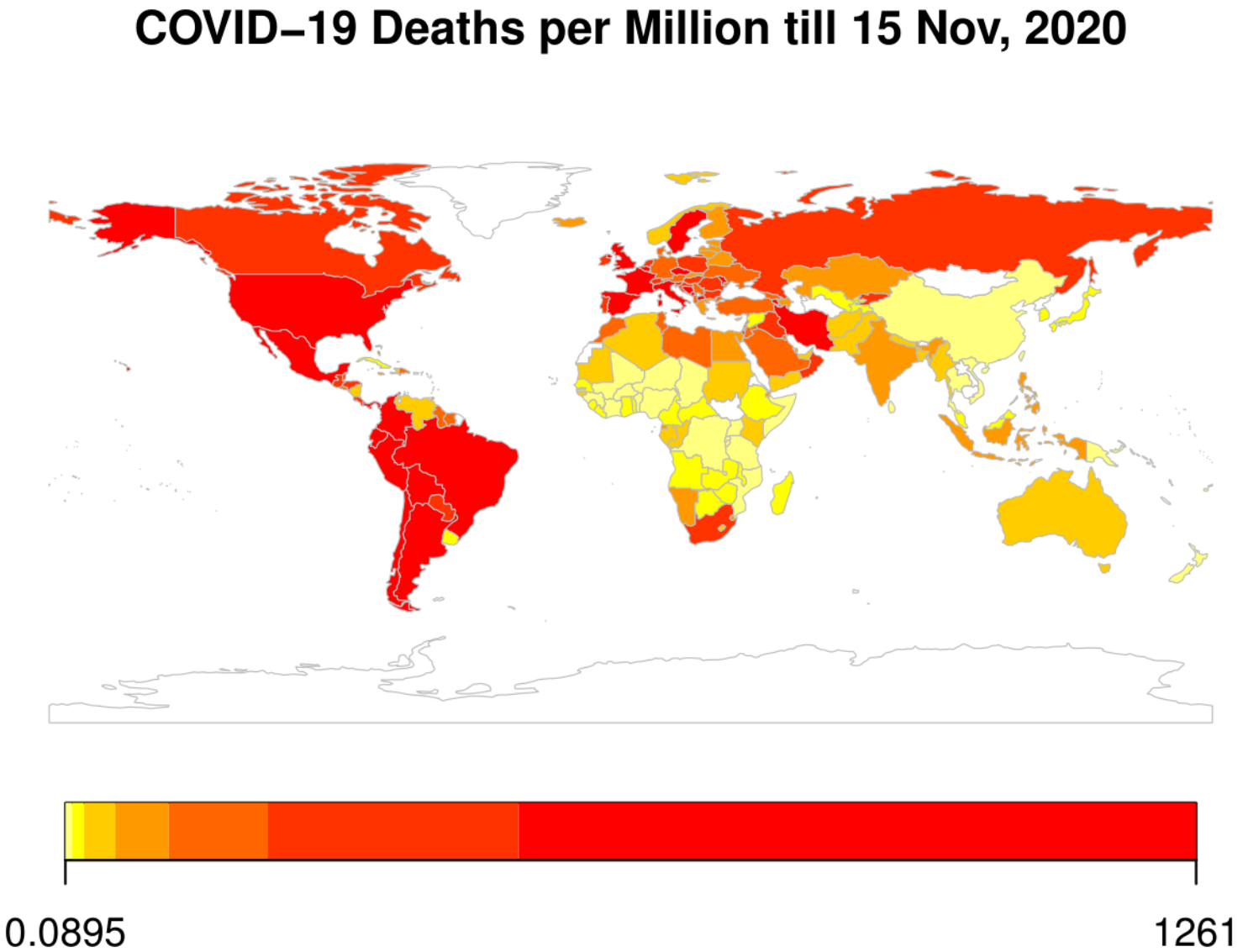
Paper Information

- This paper was published in Proceedings of National Academy of Sciences of the United States of America (PNAS) on October 12, 2020.
- Authors: Luis E. Escobar, Alvaro Molina-Cruz, and Carolina Barillas-Mury.
- DOI: https://doi.org/10.1073/pnas.2008410117
- Supplementary material (Appendix + Datasets): www.pnas.org/content /suppl/2020/07/07/2008410117.DCSupplemental

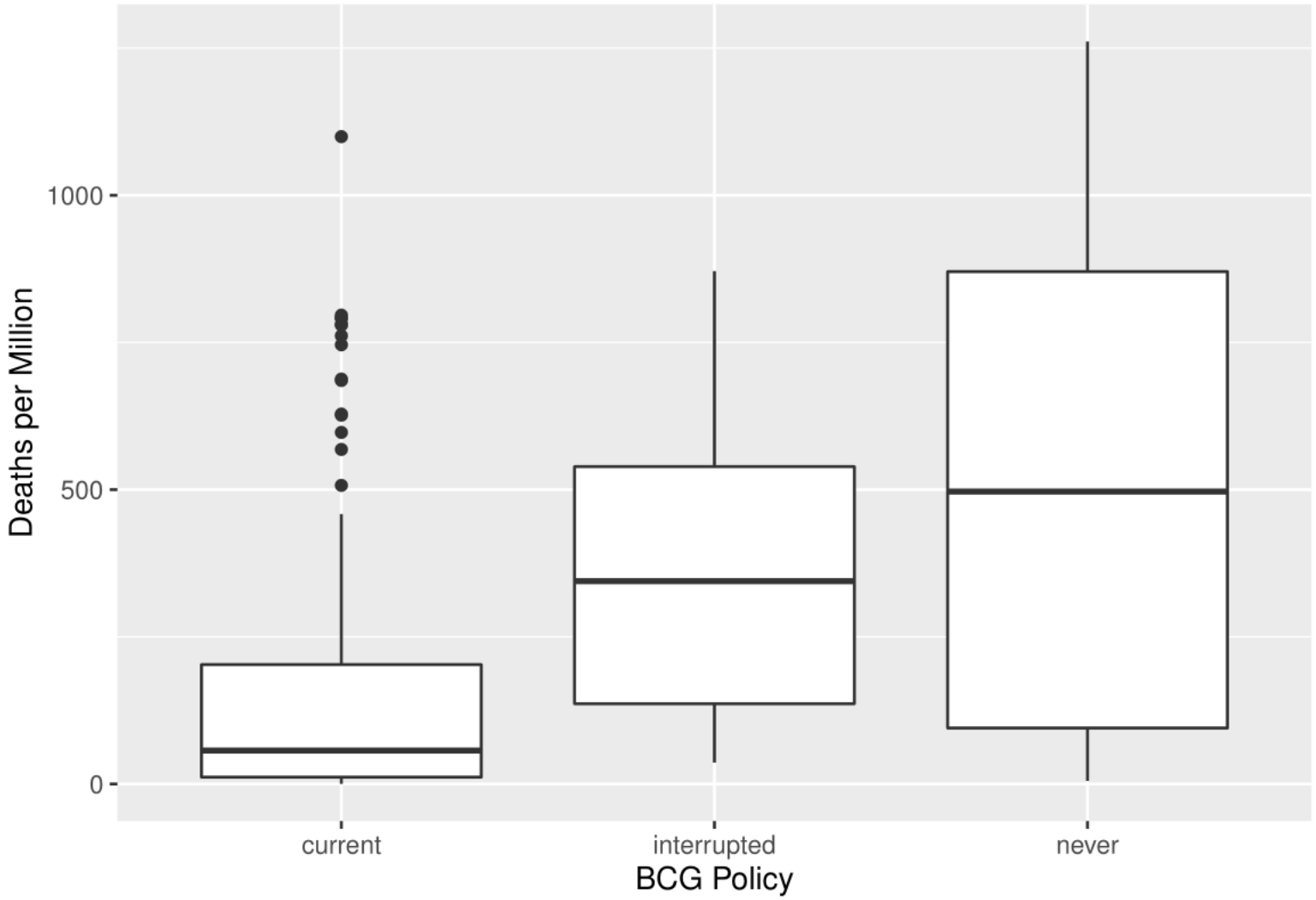
Introduction

- BCG (Bacillus Calmette–Guerin) vaccination is done in many countries to protect against TB (Tuberculosis).
- Previous studies suggested negative association between national BCG vaccination policy and COVID-19 mortality.
- Difficult to validate due to broad differences between countries such as:
  - Socioeconomic status
  - Population demographic structure
  - Percentage of population living in cities
  - Time of arrival of the pandemic
  - Criteria for testing (Number of diagnostic tests)
  - National strategies to mitigate spread of COVID-19
- This paper investigates evidence for BCG vaccine protection from severe COVID-19.

Introduction



Boxplot


Deaths per Million
1000
500
0
current
interrupted
never
BCG Policy

Methods used in original paper

- Used linear regression to conduct correlation analyses of several covariates with COVID-19 mortality. Of them - HDI, % of population >65 years, and urban %, were consistently positively associated with COVID-19 mortality.
- Used ANOVA and t-tests to assess effect of BCG vaccination policy on COVID-19 mortality. Conclusions:
  - Countries with current BCG vaccination had lower deaths as compared to countries with lack of, or interrupted BCG vaccination.
  - Percentage of BCG coverage was negatively associated with COVID-19 deaths per million.
  - Countries with higher BCG Index had significantly lower COVID-19 deaths per million.
- Filtered analysis using subset of socially similar countries.

Datasets Used

- Created new datasets by joining the ones used in the paper, along with:
  - A dataset from (gist.github.com/tadast/8827699) containing latitude and longitudes of countries, along with their ISO3 codes.
  - A time series dateset obtained from the Johns Hopkins Covid data repository on github, containing the cumulative number of deaths every day in each country, from 22nd January till 11th November (i.e. 299 days total).
- Too many observations (39550), hence aggregated above dataset to obtain:
  - cumweeklydeaths: Dataset with 5573 observations of 15 columns, giving the cumulative weekly deaths every day since the first day the pandemic hit each country (170 countries in total).
  - countryinfo: Dataset with 171 observations of 14 columns, obtained by choosing the cumulative deaths in each each country on the final observed day.

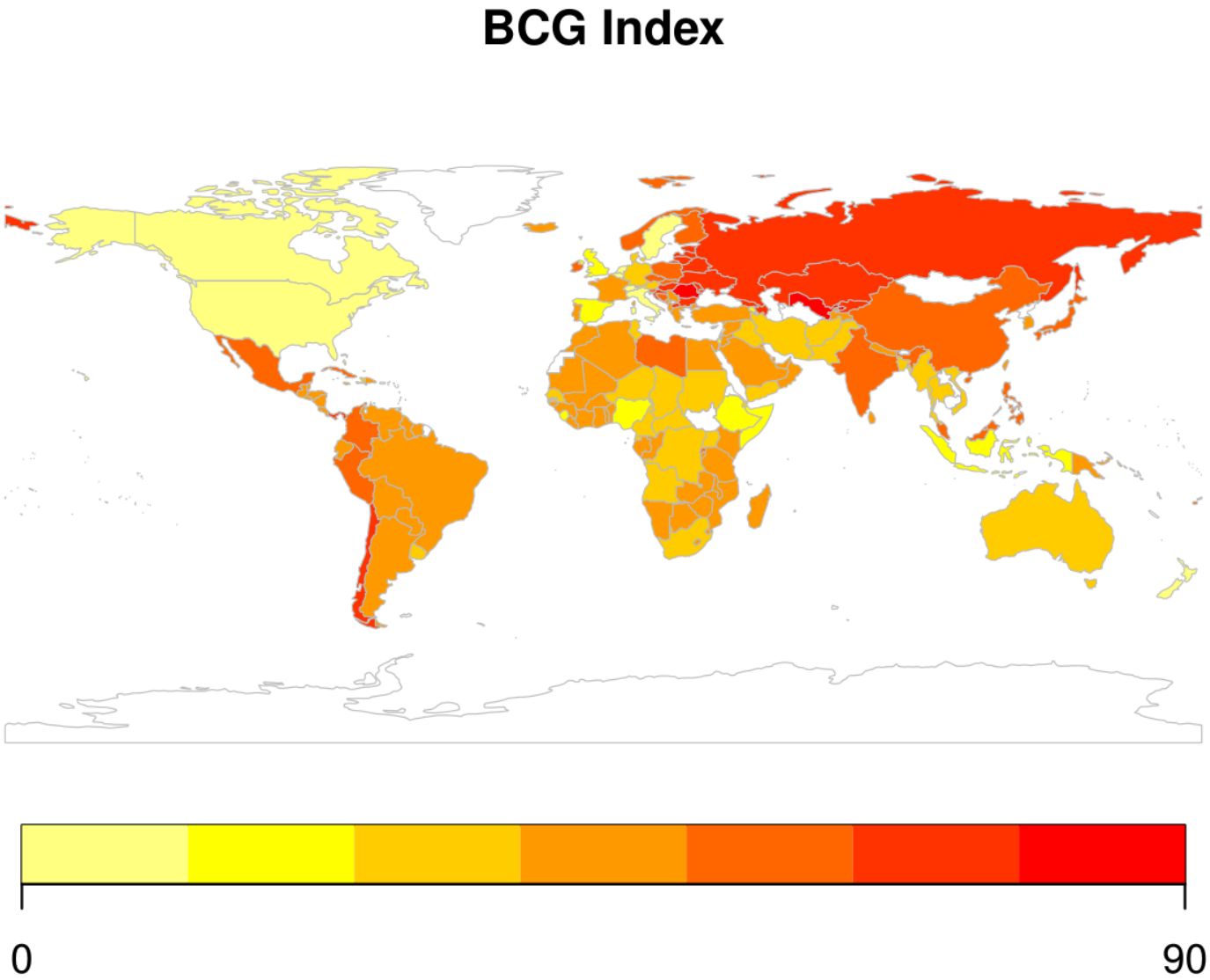
Columns of cumweeklydeaths

- Country (171 levels)
- Weeks since the first COVID-19 death in a country.
- ISO3 (Unique 3 digit identifier of every country)
- Population of country in millions (Continuous variable)
- Natural Logarithm of the population density. Log transformation was taken because this was a positive covariate, and the highest value was several orders of magnitude larger than the lowest. Original paper did not take this transformation.
- Urban fraction (Continuous, 0-1)
- Fraction of population aged 65+ (Continuous, 0-1)
- Human development index (Continuous, 0 to 1)

Columns of cumweeklydeaths

- Logarithm of deaths per million population in a (Country, Week) pair.
- Mean latitude of country.
- Mean longitude of country.
- BCG Policy (Categorical with 3 factors: current, interrupted and never)
- The range of ages of people BCG vaccine was administered on in the country.
- BCG Mean coverage (Continuous, 0-1)
- BCG Index (estimate of BCG penetration in a country) defined by the product of above two variables. Highest for Romania (89.96).

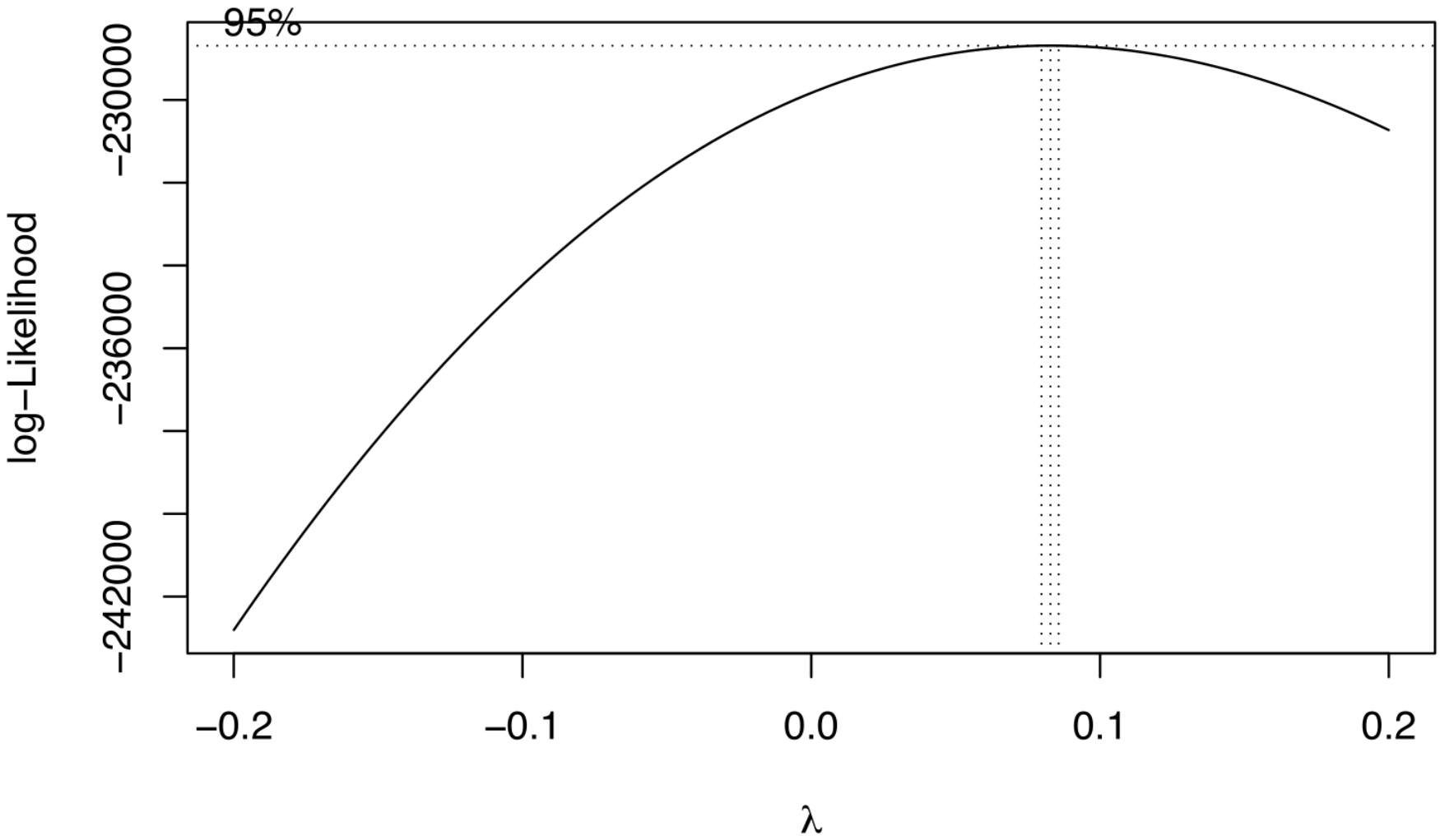
Introduction



Imputation of missing values

- Manually filled up some missing values after looking up the internet.
- For range_age_BCG, set the values for countries with never had a BCG vaccination programme, to zero, and for the rest, impute using the mean of the observed ranges of other countries, i.e. 54.45 years.
- Logistic regression of ages_65_up on the covariates HDL, log_pop_density, and urban_fraction.
- Logistic regression of BCG_mean_coverage on the covariates HDL, ages_65_up, log_pop_density, and urban_fraction.

Log transformation of response



t-test

- I took a subset of countries from the entire dataset for whom, at least 228 days had elapsed since the first COVID-19 death (130 countries total).
- Ran an 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.
- Results obtained:
  - *t:* 2.0594.
  - *df:* 10.376.
  - *p-value:* 0.06544.
- 95% CI 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.

Issues with t-test

- t-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.
- 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.

Spatial model

- We consider the following model:

$$Y_c = \beta_0 + \beta_1 HDI_c + \beta_2 log\_pop\_density_c + \beta_3 ages\_65\_up_c + \beta_4 BCG\_index_c + \beta_5 urban\_fraction_c + \epsilon + Z_c$$

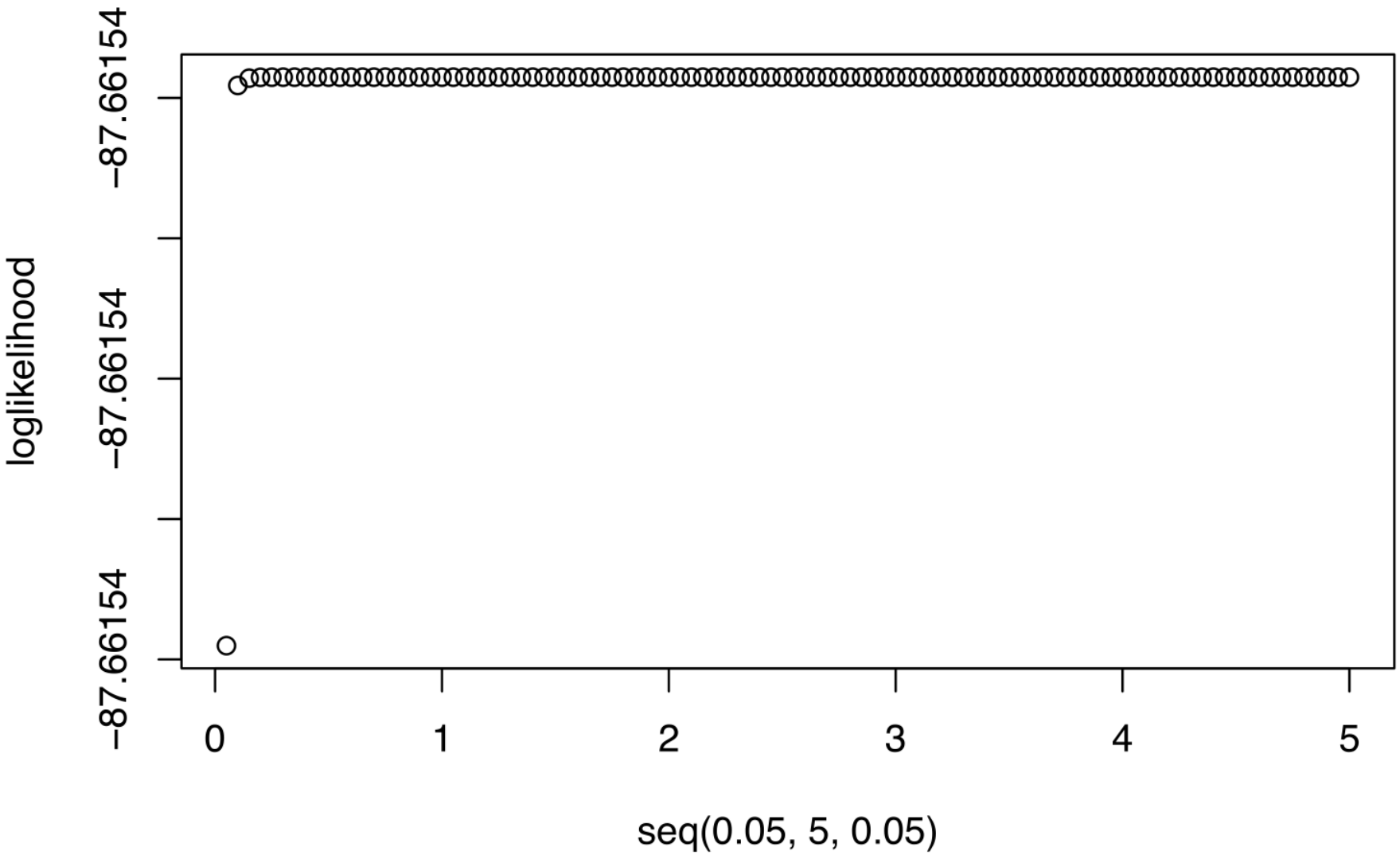
where c denotes Country, $\epsilon_c \sim N(0, \sigma_0^2)$ is random noise independent for distinct countries, and

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

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 maximized log-likelihood of the above model (using REML) fitted with the values of ρ going from 0.05 to 5 with an increment on 0.05.

Spatial model



Case $\rho \to 0$

- As we could see in the previous slide, the maximized log-likelihood remains almost constant over the entire range of ρ in consideration.
- We could simply take $\rho \to 0$, keeping $\lambda = \frac{\rho\sigma_1^2}{20037508}$ constant.
- Then, Taylor expansion would give us

$$\sigma_1^2 \exp\left(-\frac{\rho\|x_i - x_j\|}{20037508}\right) \approx \sigma_1^2 - \lambda\|x_i - x_j\|$$

- Thus, we could just consider the covariance function

$$K(x_i, x_j) = -\lambda\|x_i - x_j\| ,$$

instead of the exponential covariance function.

Spatial model

- Parameter estimates and standard errors are as follows:

Parameter	Corresponding to	Estimate	Std.Error
β_0	Intercept	4.746	2.241
β_1	HDI	0.021	1.410
β_2	log_pop_density	-0.085	0.086
β_3	ages_65_up	-6.984	3.389
β_4	BCG_index	-0.018	0.007
β_5	urban_fraction	1.465	0.791
σ_0^2	ϵ	0.626	0.165
σ_1^2	Σ	8.060	2.441

The mean of the logarithm of deaths per million decreases by 0.018 with a standard error of 0.007, for every increase in BCG index by 1.

Spatial model

- 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 5.780 on 1 degree of freedom, which gives the corresponding p-value as 0.016.
- Thus, from this analysis we get that the effect of BCG index is statistically significant.

Spatiotemporal model 1

- We now consider the following model:

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

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, and

$$Z \sim N(0, \sigma_1^2 \Sigma), \ where \ \ \Sigma[i(c_1, t_1), \ j(c_2, t_2)] = (t_1 \wedge t_2) \exp\left(-\frac{\|x_{c_1} - x_{c_2}\|}{20037508}\right)$$

is the Hadamard product of a brownian motion covariance matrix with an exponential spatial covariance matrix. $\|.\|$ is the Haversine distance as before, and I have not introduced ρ since this this model takes hours to run just once. 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.

Note that $t_1 \wedge t_2 = t_1 + t_2 - |t_1 - t_2|$. We could drop the nonstationary component $t_1 + t_2$ in this case, and simply use the covariance function $K(t_1, t_2) = -|t_1 - t_2|$.

Spatiotemporal model 1

- Parameter estimates and standard errors are as follows:

Parameter	Corresponding to	Estimate	Std.Error
β_0	Intercept	-3.230	1.090
β_1	Time (in Weeks)	0.144	0.144
β_2	HDI	2.188	0.586
β_3	log_pop_density	0.208	0.085
β_4	BCG_index	-0.016	0.003
β_5	ages_65_up	-8.067	1.948
β_6	urban_fraction	0.494	0.360
σ_0^2	ϵ	0.103	0.007
σ_1^2	Σ	1.511	0.105

The mean of the logarithm of deaths per million decreases by 0.016 with a standard error of 0.003, for every increase in BCG index by 1.

Spatiotemporal model 1

- 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 21.54 on 1 degree of freedom, which gives us a statistically very significant p-value of 3.450×10^{-6} .
- From this analysis also, we get that the effect of BCG index is statistically significant.

Spatiotemporal model 2

- Our final model is:

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

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

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

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

$$Z \sim N(0, \sigma_3^2 \Sigma_{st}),\ where\ \ \Sigma_{st}[i(c_1, t_1),\ j(c_2, t_2)] = (t_1 \wedge t_2) \exp\left(-\frac{\|x_{c_1} - x_{c_2}\|}{20037508}\right) .$$

Note that Σ_{st} is simply the hadamard product of Σ_{space} and Σ_{time}. I believe this model intuitively makes the most sense as σ_1 can be thought of as a volatility coefficient that is common for all weeks, σ_2 is common for all countries, hence their inclusion should be vital.

Code takes long time to run. More work required.

Final Thoughts

- 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.
- Clinical trials are what should generally be used to make inferences about individuals.
- Ecological Fallacy: relationships observed for groups might not necessarily hold for individuals.
- Aggregate data often easier to obtain than data on individuals, hence ecological inferences are often made. However, confounding and aggregation bias make the results unreliable.