Advanced Statistics for Genomics

Project Proposal

Jawad Chowdhury, 801135477

Topic: Statistical analysis of the influence of social, economic, and environmental factors on COVID-19 across different countries.

Introduction: In recent times, there has been numerous research ongoing based on the COVID-19 pandemic. Existing research is mostly devoted to anticipate the nature of the pandemic and explaining its different characteristics. These researches consider different social, economic, and environmental factors and try to figure out the relation between the dependent variables such as the number of infected people or the basic reproduction number with those influencing factors. In this project, I aim to analyze how these influencing factors are related among themselves, how they impact the dependent variable, and how these trends differ between different countries or regions from a statistical perspective.

Baseline work/Paper: For the statistical analysis, I am considering the following work or paper as the baseline work: "Social, economic, and environmental factors influencing the basic reproduction number of COVID-19 across countries" [1]. The main objectives of the study in Ref. [1] are-

i) to assess whether the basic reproduction number (R_0) of COVID-19 is different across countries,

ii) and to analyze the demographic, social, and environmental factors other than interventions that characterize the vulnerability to the spread of the pandemic.

Since different countries or regions have its own characteristics that expose them to the vulnerability of COVID-19 in different extent, any study that aims to measure the effectiveness of interventions across different locals should consider these statistical differences.

Dataset: The baseline work presents their data and implementation in Ref. [2]. In their study, they have considered 58 different samples (for this study it's 58 different countries). On the other hand, they have analyzed the data of these 58 countries on 7 different categories such as demographics, disease, economics, environmental, habitat, health, and social. These 7 categories form 16 different covariates to perform the statistical analysis. A glimpse of the considered dataset [3] is given below:

	country	R0_old	region_sim	days_timeseries	day_30cases	Per_reported	Log10GDP	Pop_bw20_34	logPop_tot	Pop_Urban	Urban_pop	log_Air_trans	Temp	logPercip	GHS	lo
2	Algeria	3.25	AF	16	65	33	3.606843363	26.1	7.616887011	6.43	72.05	-0.82	14.94	0.954242509	23.6	1.3
3	Argentina	1.44	SA	33	62	25	4.164110654	23.1	7.643894751	43.13	91.75	-0.42	19.41	1.827304641	58.6	1.8
4	Australia	5.77	AS	16	55	88	4.732928256	21.5	7.390967943	60.7	85.9	0.48	26.79	1.862012051	75.5	1.2
5	Austria	3.97	EU	18	51	56	4.676067775	19.7	6.944362534	21.4	58.1	0.26	-1.36	1.84416641	58.5	1.0
6	Azerbaijan	1.66	EUAS	11	71	85	3.617743461	27	6.993614012	22.92	55.34	-0.62	1.86	1.489677292	34.2	1.2
7	Belarus	2.43	EU	21	74	99	3.76055441	21.7	6.977644236	20.9	78.13	-0.59	-4.33	1.512150537	35.3	0.8
8	Belgium	2.31	EU	38	49	13	4.64561481	18.8	7.055957438	26.9	98	0.08	3.26	1.84167225	61	1.7
9	Brazil	1.85	SA	38	57	15	3.994798702	24.9	8.317716243	41.67	86.31	-0.34	25.58	2.355566448	59.7	1.6
10	Canada	1.83	NOR	40	55	17	4.653886594	20.8	7.562771728	45.9	81.4	0.4	-22.2	1.410608543	75.3	1.2
11	Chile	1.92	SA	26	59	86	4.477173763	22.9	7.266477319	25.9	87.5	-0.02	11.02	1.536178533	59.2	1.4

Research Goals: The fundamental goal of this project is to analyze the dataset from a statistical perspective and to validate if the data agrees with different underlying statistical assumptions. At this stage, the research questions or goals can be formulated as below:

(i) to analyze how these 16 different covariates are related to each other?

(ii) to explore how the dependent/target variable (which for this project is the basic reproduction number (R_0)) is related to the covariates?

(iii) since there are 58 different countries or samples, we can run a pair-wise analysis between the covariates from two different samples and check the relations between them. These can be done on a subset of all the countries. On the other hand, these data from different countries can be grouped into different continents or regions and can have a similar analysis on them.

(iv) in later phase of this project, we can formulate different models by relating the basic reproduction number on the covariates and check whether these relations differ for different countries (indicating different characteristics or trend) or they are similar in general for all the countries.

References:

[1] Kong, J. D., Tekwa, E. W., & Gignoux-Wolfsohn, S. A. (2021). Social, economic, and environmental factors influencing the basic reproduction number of COVID-19 across countries. *PloS one*, *16*(6), e0252373

[2] https://github.com/Jdkong/COVID-19

[3] https://github.com/Jdkong/COVID-19/blob/main/Data/variables.csv