Estimating the basic reproduction number R_0 of COVID-19 using a Monte Carlo method on the SIR model of differential equations for the first case peak

Contents

1 Introduction

1.1 Purpose of the Project

The COVID-19 pandemic has impacted many areas of public and private life, and once again manifested the threat of infectious diseases to human populations around the world. While posing severe challenges to public health systems, the combined efforts of health administrations, governments and citizens led to an unprecedented abundance in data on the dynamics of the spread. From a mathematical and data perspective, one might ask how this data can be translated into an effective model for the spread of infectious diseases. The promise of this is an increase in preparedness for a potentially new and similar virus, or even new mutants and strains for existing viruses like COVID-19. Additionally, such a model would enable us to quantify the impact of public health decisions, and thus facilitate evidence-based policy making. Hence, the overall purpose of the project is to use data science in order to infer characteristics of a pandemic and improve lives.
In the scope of this project, I am using the well-understood SIR model for infectious diseases (3.1) to estimate the basic reproduction number of COVID-19 (1.2) in Germany based on case and recovery data from the first case peak in late March 2020.

1.2 Background: Basic reproduction number R_0

Since the central objective of the project is the work with the so-called basic reproduction number R_0 (say: "R naught"), I shall provide a definition and description of this metric here. Citing popular literature on the subject, "[t]he basic reproduction number R_0 is the number of secondary cases which one case would produce in a completely susceptible population." (Dietz, 1993) That is, R_0 provides information on how many new infections an already infected individual is expected to cause, on average. While the idea behind this value is simple, it "depends on the duration of the infectious period, the probability of infecting a susceptible individual during one contact, and the number of new susceptible individuals contacted per unit of time." (Dietz, 1993) Hence, it becomes apparent that estimating R_0 poses not only a challenge in terms of model configuration, but especially data compilation, since many of the confounding factors are extremely difficult to determine in real-life situations. Dietz thus explains that "[t]herefore R_0 may vary considerably for different infectious diseases but also for the same disease in different populations." (Dietz, 1993) Narrowing our focus down on one country at first, and restricting our choice of data to the very first COVID-19 case peak, alleviates some of the challenges. For example, this solidifies the assumption that I am only estimating one variant of COVID-19. Nonetheless, I will have to make some strong simplifying assumptions, such as perfect susceptibility and homogeneity in the population.
After all, R_0 provides an invaluable tool in the analysis of a pandemic, since a value of R_0 greater than 1 reveals a tendency for exponential growth, while a value less than 1 indicates exponential decay.(3.1) Hence, the basic reproduction number contains plenty of information on the severity of the spread of a pandemic. Quoting Dietz again: "Because the magnitude of R_0 allows one to determine the amount of effort which is necessary either to prevent an epidemic or to eliminate an infection from a population, it is crucial to estimate R_0 for a given disease in a particular population." (Dietz, 1993) Enabling this kind of quantitative

evidence will serve the purpose described in (1.1) for is exactly our motivation.

1.3 Target Audience

There exist multiple audiences that can benefit from the insights gained by estimating the basic reproduction number or other disease-related metrics. The central target audience is clearly the group of public health officials. This group has to make decisions based on evidence and educated guesses derived from good models. For example, if a model can reliably verify that certain public health measures, such as school closures or a lockdown, posed effective in preventing the spread of a pandemic, such knowledge can be used as precedence for new epidemiological decisions in the future.
However, another equally important target audience is the general public as such, which is interested in fundamental attributes of pandemics, and how they might impact their lives. As we will see, the method used in this project allows some insight into how far and quickly a pandemic will likely spread, so sharing this information (given it stands on factual grounds) can raise general awareness of threats to the population.
Overall, this Capstone project shall serve as bottom-up approach of how epidemiological estimation can work, and hence provide incentive to explore different and more elaborate models with the computational tools available. While this project is rudimentary in its nature, its general purpose hence also targets people in academia.

1.4 A priori limitations

Before delving into the topic, I want to emphasize and disclaim some limitations of my approach.
There is no aspiration for perfect accuracy (which is a contentious aspiration in such a field anyways), and numerous simplifying assumptions should be kept in mind that are justified by the scope and time limitations of the project, but which would be imperative to consider in order to get a more realistic and accurate estimate of the basic reproduction number. Some of those simplifying assumptions include a constant population with perfect initial susceptibility and no reinfections, time-independent contact rate, homogeneous population mixing without spatial limitations, and a fixed rate of testing. (4.3)
While these assumptions are clearly quite rigid, they do not a priori prevent the obtained estimates from being useful as appropriate approximations. Also, the overarching objective of the project is the engagement with methods to estimate epidemiological measures, and not perfect accuracy.

2 Dataset

In order to estimate the basic reproduction number (1.2) in Germany for the first case peak in late March 2020, I utilize global COVID-19 data compiled by the Johns Hopkins University. (Johns Hopkins University, 2023). The data is retrieved from the official GitHub page, "CSSEGISandData"[1] maintained by the Johns Hopkins Whiting School of Engineering. In particular, the dataset comprises COVID-19 infection, recovery and death data for most of the COVID-19 pandemic, and a large set of countries. I will firstly focus on the data for Germany in the first months of the pandemic, and then extend our analysis to country data from South Africa, Peru, Russia and India. Some basic data preprocessing techniques are applied in order to convert the cumulative data into daily case numbers that can be used as a Goodness-of-Fit measure for our predictions (3.2). Despite being publicly available and curated from a prestigious academic institution, we should be aware of some inherent uncertainties in the data reporting, which can originate in a variety of causes (such as the very act of testing and reporting on the communal level).[2] A sample of the data can be found in (2),(1).

3 Method

This project uses a Monte Carlo simulation based on a system of differential equations for different subsets of the relevant data in order to simulate the case, recovery and death numbers, and then finds the parameters that best fit to the data. In order to understand the underlying workings, it is crucial to first shed light on the mathematical background of the core epidemiological model for our analysis – the SIR model.

3.1 Background: SIR model

SIR is a compartmental model that attributes members of a population to three compartments: **S**usceptible are those who can contract a disease, **I**nfected are those infected with and thus capable of transmitting the disease, and **R**emoved are those who are not susceptible any more, either due to recovery (implying immunity) or death. A system of non-linear differential equations relates the three compartments.

$$\begin{aligned} dS &= -\frac{\beta I S}{N} \\ dI &= +\frac{\beta I S}{N} - \gamma I \\ dR &= \gamma I, \end{aligned}$$

where $S + I + R = N$ is constant, i.e., I assume that no external individual gets added to any compartment during the estimation process.[3] The differential operators on the left

[1] https://github.com/CSSEGISandData

[2] Especially in the comparison of different countries, the differences in approaches of data reporting might be substantial.

[3] This is probably a reasonable assumption for the relatively short time frames employed here.

hand side of the equations imply that we are considering a change to the distribution of compartments in continuous time. In the implementation of this model, I will use a similar system to represent difference rather than differential equations in order to adapt it to the discrete real-world data. Of central importance are the two parameters β and γ, which are best understood by examining the first and third equation. The first equation describes the change of numbers in the susceptible compartment. It depends on the number of currently infected individuals I, the fraction of susceptible individuals in the entire population S/N, and the parameter β. The negative sign indicates that the particular number of individuals leaves the first compartment to migrate to the second one (i.e., it is the number of new infections in one time step). Now β is "a fixed number [...] of contacts per day that are sufficient to spread the disease." (Mathematical Association of America, 2022) This makes intuitive sense, since the number of newly infected individuals now is simply the product of all infected with the number of susceptibles they came in sufficient contact for a spread. Note that a homogeneous mixing in the population has to be assumed for this. Now the third equation states that the number of new removed indivduals (either by recovery or death) in one time step is a fraction of the individuals infected. The parameter that controls how many infected individuals are removed from the compartment in one time step is γ. Interestingly, we can thus assign a tangible meaning to $1/\gamma$ – the "average duration of infection."(Mathematical Association of America, 2022)
Given this background, the equation for dI can be understood as the number of susceptibles being added, reduced by the infected that move to the Removed compartment. Note again that the assumptions surrounding infection and removed are very binary, and do not resemble the complexity of infection and recovery processes in the real world very accurately.

To conclude this section on the mathematical underpinning of the project, I want to emphasize the meaning of $R_0 = 1$ being a cutoff, or bifurcation, value for the system above. This can most easily[4] be seen by considering the second equation. One would agree that the spread of the disease is accelerating if the number of infected individuals increases, i.e., if $dI > 0$, and likewise that the disease is dying out for $dI < 0$, with stagnation achieved at $dI = 0$. So assume the disease is neither accelerating nor dying out. Then this implies

$$+\frac{\beta IS}{N} - \gamma I = 0 \Leftrightarrow \frac{\beta IS}{N} > \gamma I \Leftrightarrow \frac{\beta}{\gamma} = \frac{N}{S}.$$

Now one central assumption to the model was that initially, the entire population is susceptible, so we can write $\frac{N}{S} \approx 1$, which clearly holds true for the initial phase of the spread. So defining $R_0 = \frac{\beta}{\gamma}$, we imprint the cutoff value $\frac{\beta}{\gamma} = R_0 = 1$ between exponential spread and decay into the basic reproduction number, which gives us a very useful tool to estimate the basic reproduction number.

3.2 Monte Carlo simulations

Having obtained the relation $R_0 = \frac{\beta}{\gamma}$, I set up a Monte Carlo simulation to find the best fit of the parameters of the SIR model to the given data, and then calculate R_0 based on

[4] yet not comprehensively, since we would have to consider the system in more depth

those best estimates. In order to do so, I leverage an implementation of a similar idea from Sherry Towers, who has developed a very useful method for the estimation of influenza data (Towers, 2013). As mentioned earlier, I use a system of difference equations rather than differential equations to account for the discrete nature of the data. Also, I adapt the approach developed by Towers to estimate R_0 assuming a point of completely agnosticism towards both parameters of the model. For example, if we knew that the average infectious period is 3 days, we could set $\gamma = 1/3$, and simulate using this information. However, I want to display a bottom-up method in which we have no knowledge about the contact and removal rate whatsoever. In order to do so, I proceed as follows: For every country (starting with Germany), I determine the first case peak. This will be the central point for an interval containing 11 days (5 days prior and 5 days post the peak, plus the peak day itself). For the 11 days from start of the pandemic,[5] I run a Monte Carlo simulation with 1500 iterations, supplying ranges for β and γ and the population size of the country. In particular, I chose the ranges to be $(1/16, 3)$ for β and $(1/8, 1/2)$ for γ. In each iteration, a pair (β, γ) is sampled from the specified ranges, and provided to the system of difference equations in order to simulate the spread of the pandemic. This results in estimates for infected and removed individuals over the respective time frame. Such a fit can be seen in 3

Since those estimates need a metric for comparison to the actual data, I use a Poisson Negative Log-Likelihood function that calculates how well the data matches both the actual case and removed data. It should be pointed out that the real-world removed data is the sum of death and recovery data obtained from the Johns Hopkins datasets. The best percent of estimates (15 out of 1500) is extracted for each time step (i.e., I store the pair of parameters in a data frame for later use). After all runs are finished, I thus have 165 pairs of parameters that resemble the respective best fits for each time frame. In the following, I can do analysis on those pairs, and in particular, get an estimate on R_0.

A three-dimensional plot of the error function with respect to β and γ can be found in (4).

[5]The start of the pandemic is defined for each country individually as the date of the first reported case.

4 Results and Interpretation

4.1 Estimating R_0 in Germany

I apply two different but closely related ways to estimate the basic reproduction per country, and then use a third (external) method to validate the model predictions. Having obtained the pairs of parameters, I firstly calculate the 165 ratios of the pairs, and then simply determine the median. This yields a final estimate on R_0 for Germany of 2.044, with a 90 % confidence interval $(1.566, 2.936)$.(5) One should note that the confidence interval is extremely large, but clearly the estimate is significantly greater than 1.
A second method starts with the 165 pairs, and projects them onto the (β, γ)-plane. This is indeed a projection, since the error function before yielded a three-dimensional plot, so one can think of this as cutting the graph in (4) at the bottom (to obtain the best fits), and projecting the pairs on the plane. It is then straightforward to run a regression on the points. Unfortunately, the regression line missed the origin, so the slope of the regression line is not helpful in determining the ratio of the parameters. The very fact that the y-intercept was nonzero, however, hints at the fact that my estimation method is not stable, but can produce varying estimates for very similar Goodness-of-Fit statistics. However, the estimates are all in a reasonable range. Now in order to actually obtain a value from the data, I thus calculate not the standard linear regression, but regress a line through the origin.[6] Then the slope of the regression line is the desired R_0 estimate. This yields $R_0 = 1.809$, with a confidence interval $(1.760, 1.857)$.(6) Hence, the second estimation methods also yields $R_0 \approx 2$, which is again significantly greater than 1 – a reasonable result.
To verify my own results with a pre-made estimation package on the particular data I have used, I loaded the R_0 package from R and supplied my data. Utilizing the Exponential Growth method from the package, I obtained the estimate $R_0 = 2.49$, with a confidence interval of $(2.470, 2.510)$. However, it should be noted that this package needs information on the infectious period, so it is not completely comparable to my approach, and strongly depended on which information is actually supplied. In the grand picture, this external estimate validates the magnitude of my estimates. Comparing to values from epidemiological literature using different methods, I can attest a reasonable similarity of my estimates.(Prada, J.P., Maag, L.E., Siegmund, L. et al., 2022)

4.2 Results for other countries: South Africa, India, Russia and Peru

Subsequently, I extended my analysis to four other countries, in particular, South Africa, India, Russia and Peru. While I was able to demonstrate that my method is reasonably capable of producing an estimate for Germany, this case was unique among the five countries, since the first COVID-19 case peak almost coincided with the first official lockdown announcement.(9) Now since the estimate obtained by the Monte Carlo method is expected to reflect some sort of changes for different policies (assuming all other factors are held constant), the treatment of the additional countries in the light of their first peak was of particular interest,

[6] This is somewhat similar to finding the best one-dimensional subspace to the data.

since all these countries implemented social distancing measures long before the first peak occurred. (10)-(13)
A plot visualizing the results for those countries can be found in (14), but it is worth pointing out the central finding here. For each of the four lockdown-impacted countries, the estimate for the basic reproduction number was remarkably lower using either method, with some estimates even being close to 1.(1) This strongly suggests that the lockdowns had a significant impact on the prevention of the spread of the disease. It should be noted that due to the impact of the lockdown, I did not actually estimate the exact same metric as in Germany (the underlying assumptions differ with respect to the policy environment). So even if the label R_0 should not be imposed on both cases without context, it is reasonable to claim a good level of commensurability between the estimates, since the method used to obtain the estimates was identical. Granted that all other factors are relatively stable and comparable between all countries, and observing that the metric for the lockdown-impacted countries is comparably lower, I can thus claim that the data strongly suggests that lockdowns are an effective way to contain the spread of a pandemic. Especially South Africa is very interesting, since it has comparable population and infection/recovery data on the first glance. The results for South Africa can be found in (7) and (8). Now the general finding of a lockdown reducing the spread of the pandemic is certainly expected to a certain degree, but the model gives us a way to quantify the effectiveness in terms of a simple measure, and thus poses a valuable tool to assess effectiveness of public health measures.[7] Lastly, it should be noted that upon investigating the linear regressions on the four other countries, they show far lower y-intercepts than the Germany data, in relative numbers.(cf. 8) Thus, the treatment of the four countries reveals some indication that the method might work better in lower ranges of the basic reproduction number, which would be an interesting object of further study.

4.3 Limitations revisited

We have seen that R_0 is a very complex measure, determined by many different factors such as infectious period and contact rate. In our estimates, the influencing factors included the time frame of consideration, public policy measures, and various assumptions including complete susceptibility at the start of the pandemic as well as perfectly homogeneous population mixing. Additionally, the time frame of choice, i.e., the subset of the data, impacts our estimates.
So while it is nearly infeasible to obtain a *true* estimate for the basic reproduction number, this project addressed finding a *good* estimate by a bottom-up approach agnostic to previous knowledge of infectious period and contact rate. The issue of choosing a representative time frame was handled by subsetting the data around the peak, and applying averaging methods for the obtained results. This is especially relevant in the light of different mutants and strains dominating the epidemiological process, since over the first peak, we can safely assume one variant of COVID-19 to dominate throughout. Additionally, it addresses the problem of not allowing reinfections, since a reinfection during such a short period of time is highly unlikely. Lastly, the impact of public policies were discussed using the significant difference

[7] Indeed, the effectiveness of a lockdown is very intuitive and has been shown in numerous other and far more elaborate publications – thus, the fact that the Monte Carlo method reproduced this finding is a good gauge for the validity of my bottom-up approach.

in estimates for countries impacted differently by lockdowns and social distancing measures. Overall, I am confident that the general public as target audience would be able to benefit from the project, since it shows how to grasp a difficult estimation problem in epidemiology with relatively simple ideas, and produces straightforward and nicely interpretable results. The target audience of academics will likely not benefit from this project, since far more elaborate methods already exist. However, I think it is appreciable to contribute to the discourse around how to define appropriate metrics to measure diseases. Lastly, this implies that the most central target audience – public health officials – will probably only benefit from the results of those most advanced such projects. Still, behind every grand research stands a long history of trial and error, and different ideas being explored and tested.

5 Conclusion

5.1 Outlook

In future work, it might be worthwhile to consider time-dependent parameter, and to weaken the assumption of perfectly homogeneous mixing. Another factor to consider would be the testing rate, since testing more frequently correlates positively with reporting a higher number of cases naturally. Alos, using the presented approach, further information on infectious periods from clinical studies could directly be incorporated to refine the estimates. In my two-parameter estimation, however, it is difficult to infer further characteristics of infection time and contact rate, since the spread was wide (I was mainly interested in the ratio). So even though the model is not among the most complex, it is not entirely interpretable on the structural level of β and γ. For different purposes and higher accuracy, my model could serve as basis for more complex and versatile models. One should thereby always keep the common tradeoff between complexity and explainability for those models in mind – adding more parameters will likely make it more difficult to interpret the results adequately. The choice of correct model hence must be in good accordance with the objectives of the target audience of the project, and very natural in this context a rather explainable model seems desirable (granted that accuracy does not suffer to much). This is because policy-makers are primarily interested in why exactly the numbers behave as they do, e.g., in order to make decisions on the necessity of a mask mandate or so.
Furthermore, in the context of vaccinations, it is of immense interest to weaken the assumption of complete susceptibility, and examine the impact of a vaccination and immunity (as well as the risk of reinfection) on the course of a pandemic. There have been various studies on this (Bai, Fan and Brauer, Fred, 2021), and they often use more elaborate versions of the SIR compartments (or partitions of the existing SIR model) in order to gain further traction on the estimates.

5.2 Conclusion

The COVID-19 pandemic attested the increased importance of epidemiological measures that can be used for inference of the disease spread, but also for straight-forward and understandable public communication tools. Estimating metrics of infectious diseases will always remain a challenging yet imperative task in combating outbreaks, and that an engagement of as many people in the discussion around appropriate metrics and estimates is needed to mitigate the immense threat of infectious diseases to human populations around the world. Thus, whilst the estimation method provided in this project is rudimentary, it contributes to the on-going discussion by showing how to engage with some of the inherent questions on the topic. In particular, I showed that the from-scratch SIR estimation approach provides a valid model to estimate R_0, and I demonstrated that a lockdown changes the spread of an infectious disease drastically.
I may thus conclude that I completed the goal of working towards the usage data science on real-world data to infer characteristics of a pandemic and improve lives through evidence-based decision making.

6 Figures and Tables

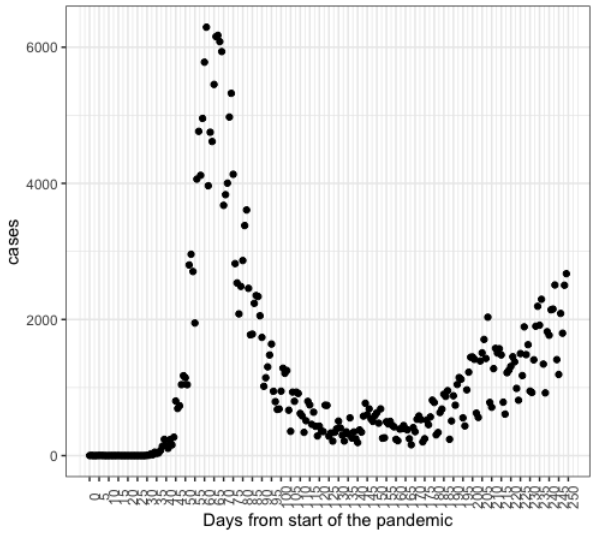

Figure 1: COVID-19 Case Data Germany

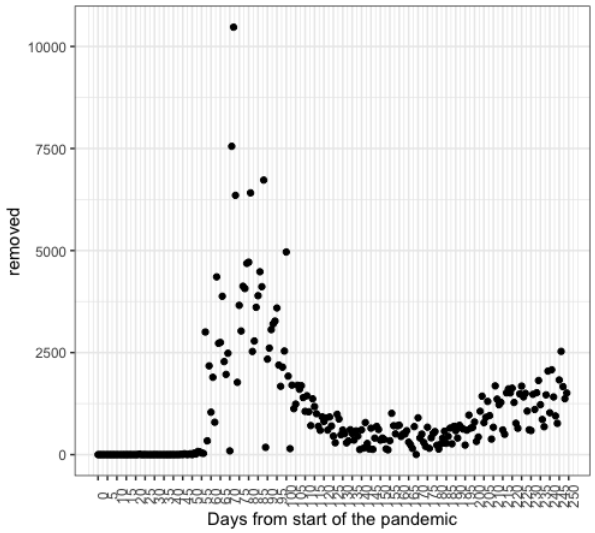

Figure 2: COVID-19 Removal Data Germany

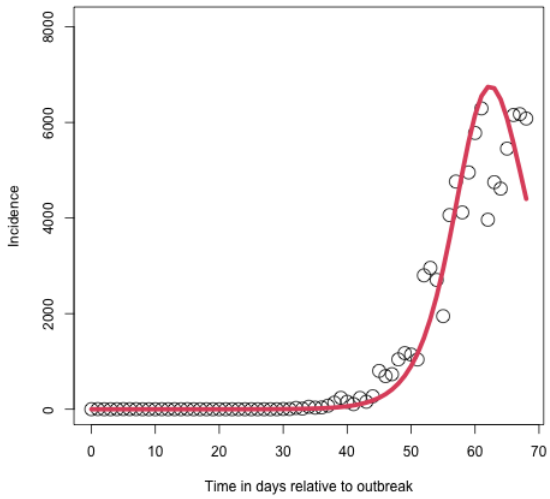

Figure 3: Model fit to incidence data in Germany

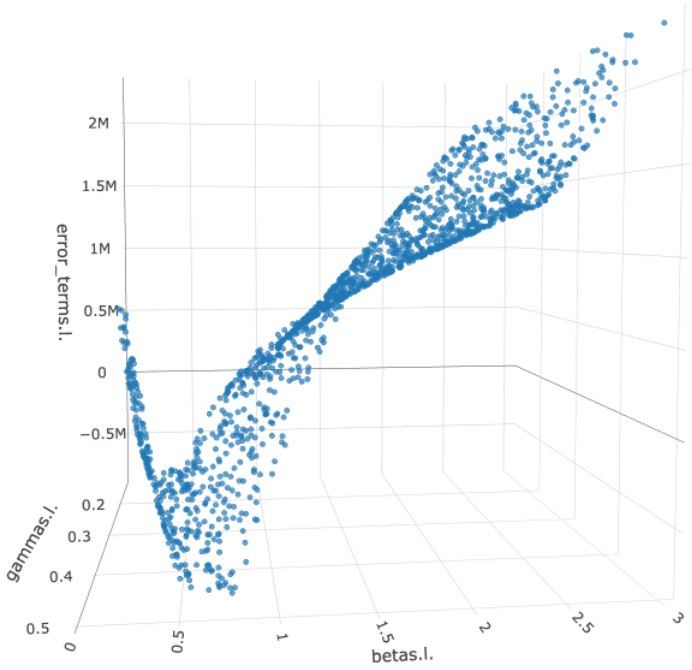

Figure 4: Error function in 3D

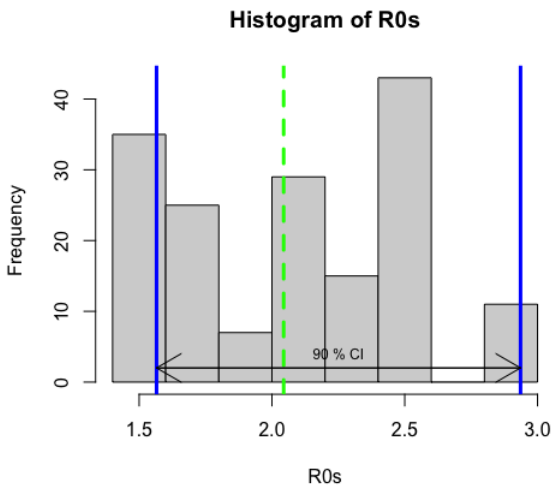

Figure 5: R_0 estimation (Germany) using percentiles (median)

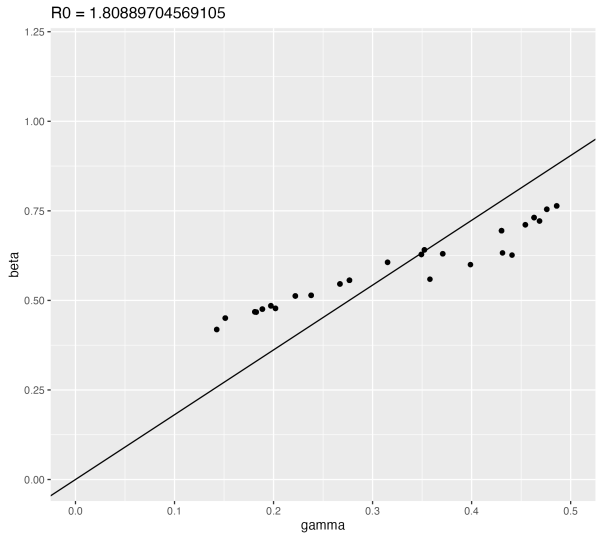

Figure 6: R_0 estimation (Germany) using regression through origin

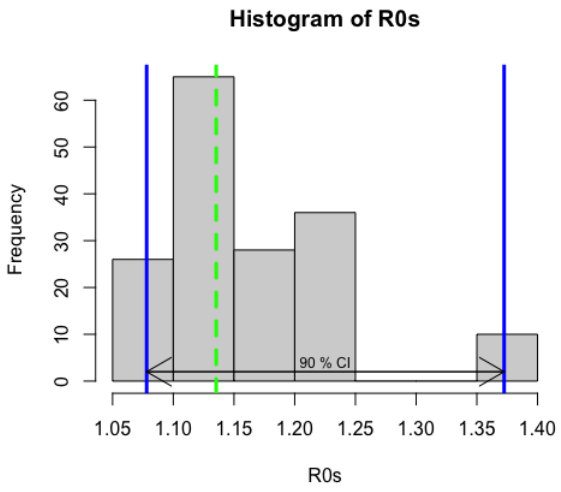

Figure 7: R_0 estimation (South Africa) using percentiles (median)

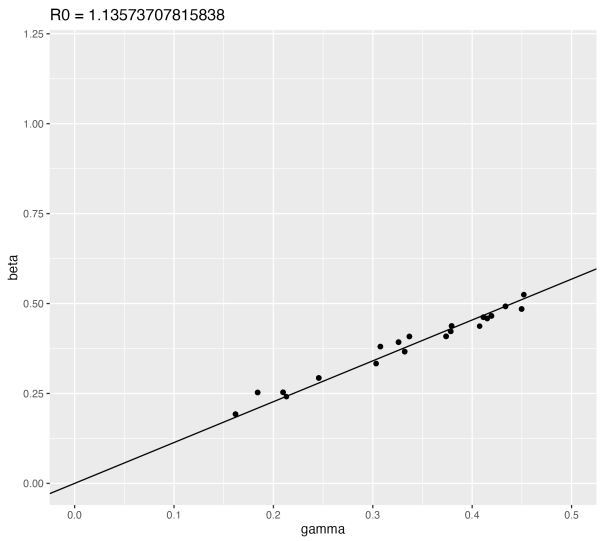

Figure 8: R_0 estimation (South Africa) using regression through origin

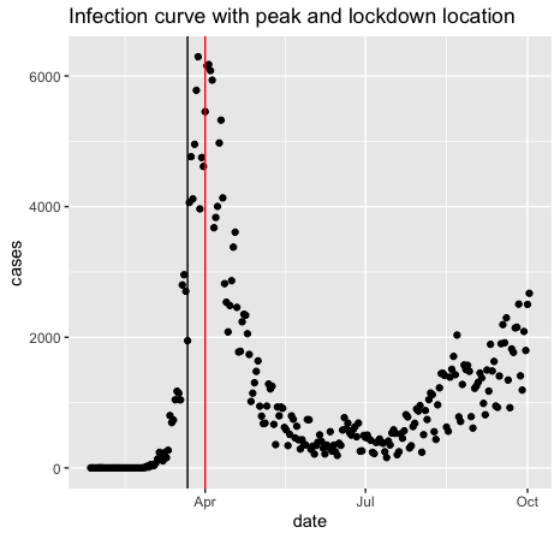

Figure 9: Germany case peak (red) and lockdown announcement (black)

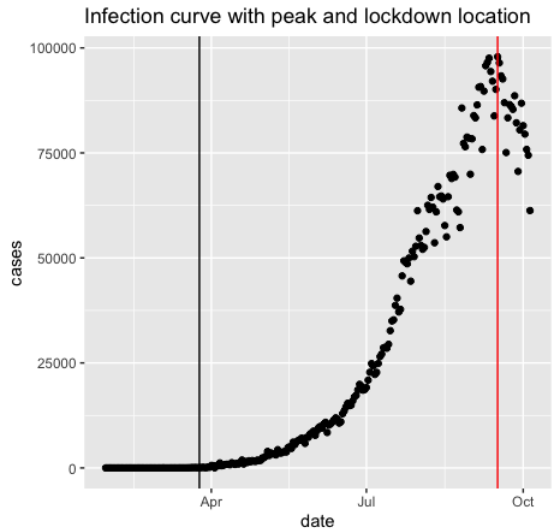

Figure 10: India case peak (red) and lockdown announcement (black)

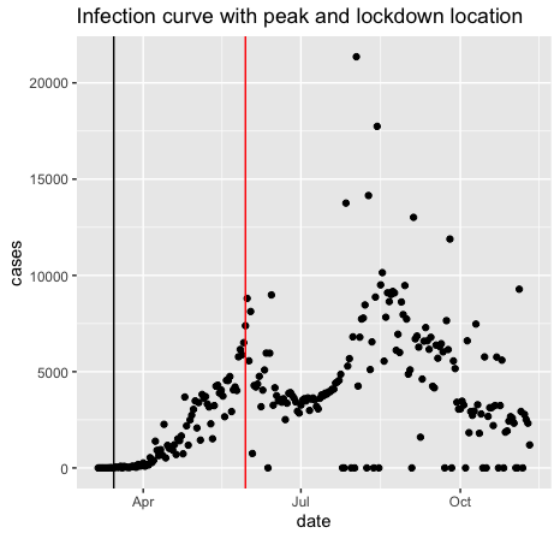

Figure 11: Peru case peak (red) and lockdown announcement (black)

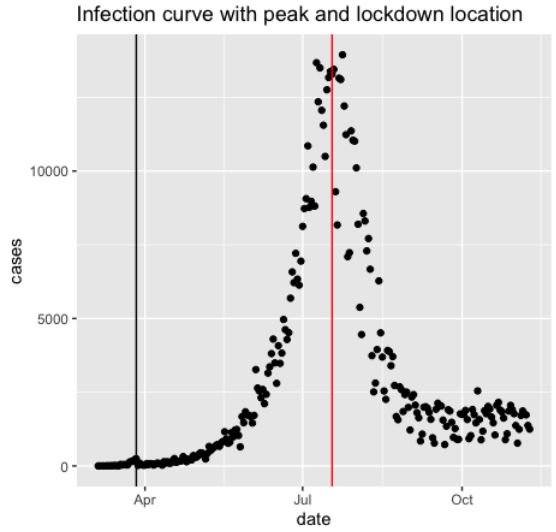

Figure 12: South Africa case peak (red) and lockdown announcement (black)

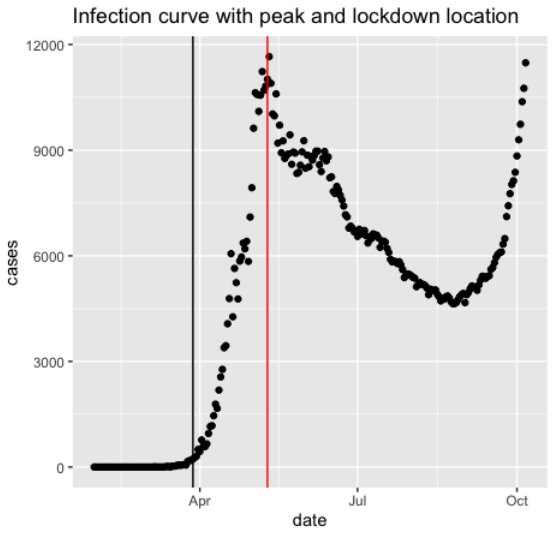

Figure 13: Russia case peak (red) and lockdown announcement (black)

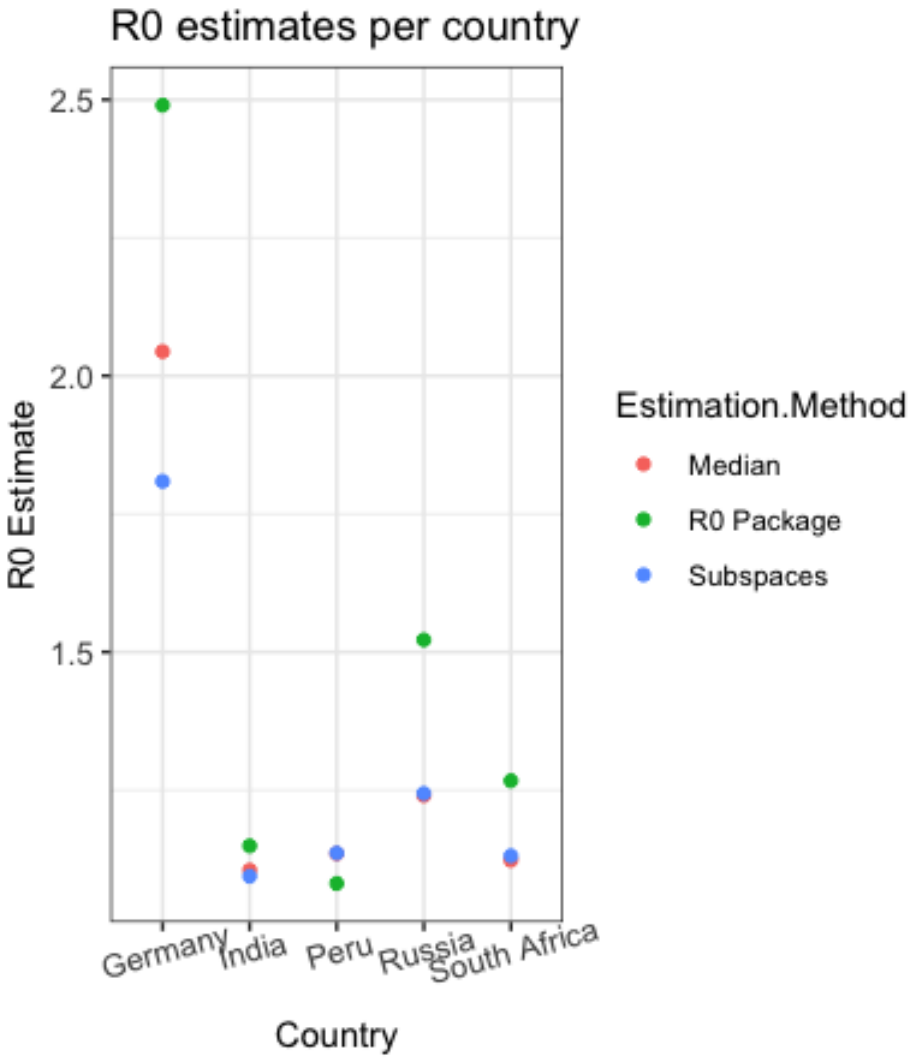

Figure 14: R_0 estimates per country

Country	Median (90% C.I.)	Regression (95% C.I.)	R_0 Package (95% C.I.)
Germany	2.044 (1.566,2.936)	1.809 (1.760,1.857)	2.490 (2.470,2.510)
Peru	1.105 (1.038, 1.200)	1.094 (1.090, 1.100)	1.149 (1.149,1.149)
South Africa	1.123 (1.078,1.236)	1.131 (1.124,1.137)	1.267 (1.266,1.268)
Russia	1.135 (1.078,1.372)	1.136 (1.129,1.143)	1.081 (1.081,1.082)
India	1.240 (1.163,1.540)	1.244(1.230,1.258)	1.522 (1.520,1.523)

Table 1: R_0 estimates per country

7 Code and Acknowledgements

The project code and comprehensive results can be found on my GitHub Repository available at https://github.com/timneumann1/DataScience_Capstone.

I would like to thank Professor Lo for the support throughout the semester, and Tom Cooklin for many inputs that helped me advance my project. Also, I would like to thank Sherry Towers for the inspiration on the Monte Carlo approach employed in this project.

References

Dietz, K. (1993). The estimation of the basic reproduction number for infectious diseases. *Statistical Methods in Medical Research*, *2*(1), 23–41. https://doi.org/10.1177/096228029300200103

Towers, S. (2013). Fitting the parameters of an sir model to influenza data using least squares and the graphical monte carlo method. https://sherrytowers.com/2013/01/29/neiu-lecture-vi-fitting-the-parameters-of-an-sir-model-to-influenza-data/#least

Bai, Fan and Brauer, Fred. (2021). The effect of face mask use on covid-19 models. *Epidemiologia*, *2*(1), 75–83. https://doi.org/10.3390/epidemiologia2010007

Mathematical Association of America. (2022). The sir model for spread of disease - the differential equation model. https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model

Prada, J.P., Maag, L.E., Siegmund, L. et al. (2022). Estimation of r0 for the spread of sars-cov-2 in germany from excess mortality. *Sci Rep*, *12*(17221). https://doi.org/10.1038/s41598-022-22101-7

Johns Hopkins University. (2023). Cssegisanddata covid-19 time series. https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series