Data Science and Information Technologies M.Sc.

Specialization: Biomedical Data Science - Bioinformatics

Machine Learning in Computational Biology – 2nd semester

Nikolas Kalavros (DS2190008)

6/4/2020

Modelling the spread of Coronavirus using polynomial models

Introduction:

It has been over ten years since the World Health Organization last declared a disease, the last notable case being the H1N1 influenza virus outbreak that took place from January of 2009 to August 2010[1,2]. A decade later, on January 30 2020, WHO declared a new viral outbreak as a "public-health emergency of international concern". Less than two months later, on 11 March 2020, it was declared a pandemic[3]. The origin of the outbreak is placed in the Wuhan province in China[4].

The virus was named SARS-CoV-2, previously referred to as 2019-nCoV and it is a β Coronavirus, specifically 2B with over 70% genetic similarity to SARS-nCoV[5]. Its genome has been since then fully sequenced and there exist many sequence examples from viral samples, collected from patients all around the world[6]. This has allowed scientists to conduct various analyses, yielding novel insight into the virus' mechanism of action, targets and potential therapeutic agents[7–9]. Aside from sequence-based studies, there have been many structural based approaches, in order to classify the virus, understand its mechanisms of action and ultimately identify therapeutic agents to neutralize the infection[10,11]. Finally, many researchers are opening up new avenues to combat the disease by examining the body's immune response[12–14].

One aspect of this outbreak that is underexamined is the social response to it. A thorough critique on this subject is beyond the scope of this assignment, however, it should be noted that the virus has been demonstrated to have occurred naturally[15]. Similar viruses have been identified before which might also indicate its origins[16–19]. Further research is sure to be carried out in order to pinpoint the virus' origins more accurately.

As for the infection itself, the virus is spread during close contact and through small droplets[4]. The virus seems to have various half-lives, depending on the surface involved, based on some experiments that have been recently published[20]. As the new coronavirus is also a respiratory disease, it can also be transmitted through gas clouds and so many countries have enforced a policy of social distancing to combat the infection[21,22].

Once infected, many people are asymptomatic, though the percentages seem to vary[23,24]. Of those exhibiting symptoms, the most common ones are fever and dry cough. There are also various other less frequent symptoms, such as headaches, production of phlegm and loss of the senses of taste and smell[4,25,26]. According to WHO, 1 in 6 people become seriously ill and some cases proceed to viral pneumonia, organ failure and ultimately death[5]. As for the mortality rate, it varies wildly and depends on age and socioeconomic factors[27,28].

It is evident that the pandemic is far from over and that measures need to be taken to curb the rate of infection. Computational studies are indispensable both in understanding the virus itself at a biological level and to model the infection at an epidemiological one. One very interesting attempt that should be highlighted is the Folding@Home initiative, which is a "distributed supercomputer" that performs Molecular Dynamics (MD) and running simulations on thousands of computers and phones worldwide. They have recently launched a new initiative in order to perform large-scale MD for the SARS-CoV-2 spike protein[29]. Their website can easily be accessed through this link: https://foldingathome.org/.

In this assignment, an attempt will be made to model the spread of the disease, based on a very comprehensive dataset, provided by the Johns Hopkins University (JHU) Center for Systems Science and Engineering (CSSE)[30].

Materials & Methods:

The dataset:

As mentioned previously, the JHU dataset will be utilized in order to produce the optimal model for the virus' growth curve. This dataset consists of three main files that will be utilized. All files are constructed in a similar way and provide nation-wide data on new confirmed cases of SARS-CoV-2, deaths by the disease and the number of patients recovered. The data is automatically updated daily, with a new column being added every day, representing the new cases, deaths and recoveries per location. The data spans from the 22nd of January 2020 to the current date. In the form that the assignment is handed in, the final date used is the 6th of April. However, the code can be rerun to create updated models, based on new data. Archived data from previous dates exists, but is incomplete and was therefore not employed in model creation.

Modelling the epidemic:

As per the assignment's instructions, a parametric family of models will be utilized, the polynomial family. More specifically, polynomials of degree from 2 to 9 will be created, in order to best characterize the growth of the virus, while avoiding overfitting. While models will be constructed for many countries, models will be displayed only for a select few: China, Italy, USA, Spain and Greece in the interest of space. Models for more countries can be also constructed using the accompanying Jupyter Notebook. In order to assess the models' performance, the Bayesian Information Criterion and the Akaike Information Criterion will both be used. This is also a parameter in the code, subject to change. Another aim is to minimize the test error.

In order to create an unbiased dataset, a date X parameter was used, which in this case was the 20th of March 2020, but is subject to change in the accompanying notebook. Following the assignment's suggestions, the data was split randomly into a training fold (66%) and a testing fold (33%). The days were converted from the MM/DD/YYYY format to integer format, in order to create integers that were easier to incorporate in the model. Subsequently, models were created for the number of infected cases, deaths by the disease and patients recovered as well.

Mathematical Background:

The polynomial family of parametric models encompasses all models of the form:

$$y_n = a_n * x^n + a_{n-1} * x^{n-1} + a_{n-2} * x^{n-2} + \cdots + a_1 * x + a_0$$

While this family of models can approximate non-monotonic functions, as well as exponential and oscillating functions, curve fitting with polynomial parametric models often overfit, are unable to extrapolate and are usually unstable. As such, a selection strategy is needed in order to identify suitable ones for this task. As noted previously, degrees from 2 to 9 are used.

The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are computed using the following formulas:

$$AIC = -2 * \log(likelihood) + 2 * parameters$$

$$BIC = \ln(datapoints) * parameters - 2 * \ln(\text{likelihood})$$

Under the assumptions that the errors are normally distributed and since polynomial regression is an extended case of least squares regression where features are expanded, the likelihood function can be replaced with the residual sum of squares, divided by the sample size. In the code, that simplification is used in order to provide an easier method to compute the two criteria. These two criteria have the distinct advantage of taking the number of parameters into account, thus providing another method by which to avoid overparametrization and overfitting.

Solving the Least Squares Regression Problem:

Given a feature matrix X of dimensions m x n, where m is the number of examples and n is the number of features and a target vector Y of dimensions m x 1, the ordinary least squares method provides a closed form expression in order to calculate the weights vector Θ, of dimensions n x 1. The formula is provided below:

$$\Theta = (X^T * X)^{-1} * X^T * Y$$

Tikhonov Regularization:

A method to regularize the parameter vector Θ and reduce the values of parameters, based on a value λ. The higher λ is, the more bias is introduced and variance is reduced.

$$\Theta = (X^T * X + \lambda * I)^{-1} * X^T * Y$$

Features used:

First, the data was split according to the aforementioned schema. For all 3 different datasets (confirmed cases, deaths, recovered patients), 4 different categories of models were constructed:

1. A model that only considers only the total number of cases (per condition) up to that day and attempts to predict the number of cases in the next day (per condition).

2. A model that considers the total number of cases (per condition) up to that day, as well as the day itself, which is represented as a repeat vector of dimension 1 x n, with the repeating element being the date, converted into integer format. This kind of model aims to account for specific trends and perhaps approximate measures taken by each nation to flatten the curve. Similarly to the previous one, it uses this data to attempt to predict the number of cases in the next day (per condition).

3. A model that considers the total number of cases, but of all three conditions. For example, in order to predict the number of confirmed cases at the 5^{th} of March, the total number of confirmed cases, deaths and recoveries at the 4^{th} of March will be used. This kind of model aims to account for the gap between the epidemiological insight that can be gleamed by only looking at the data and the actual prevalence of the virus. It should be noted that in actual epidemiological studies, the death rate is oftentimes a better indicator of the number of cases in the population than the number of confirmed cases. In the same vein with the previous model, this one attempts to predict the number of cases in the next day (per condition).

4. A model that considers the total number of cases (per condition) for that specific date and the 3 previous ones. For example, in order to predict the number of confirmed cases at the 5^{th} of March, the total number of confirmed cases and recoveries from the 1^{st} up to the 4^{th} of March will be used. This model aims to account for slight variation that might not be of a large enough scale to be captured by a model that only takes into account the previous day's cases. The model attempts to predict the number of cases in the next day (per condition).

Results:

Effectiveness of feature augmentation:

In all cases that were tested through the code, which were namely the optimal model never contained the augmented features. Those include the usage of the previous three days when performing predictions with the model, the date, when converted to integer format and the use of the recovered patients and deaths by the disease. This phenomenon can perhaps be attributed to the fact that the dataset was quite small, so any features which were unnecessary quickly led to large amounts of overfitting.

The degree of the model:

Strangely, in all cases, there was no optimal model with a degree above 2. That means that 2 features and one intercept were enough to sufficiently capture the curves. Even in countries with more complex progressions, such as China (plateau), Taiwan (mostly linear and contained) and South Korea (beginning to plateau), the optimal model still had degree two. It should however be reported that there were models of higher degree that had lower RMSE, however, they were usually burdened with many more features and as such, were pruned when using the Bayesian Information Criterion. It is suspected that the size of the dataset is the reason the degree was always so small.

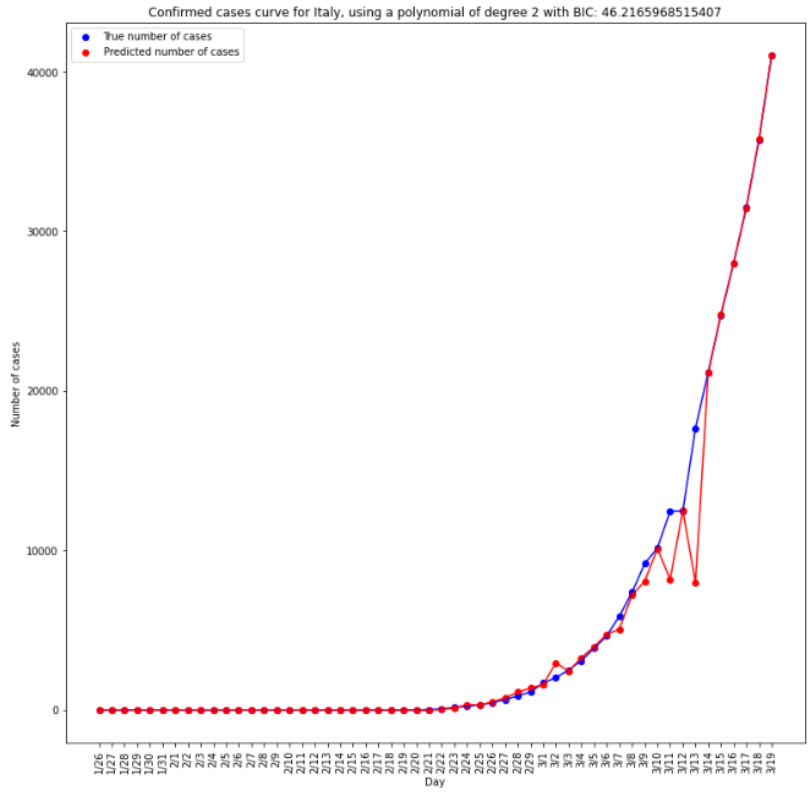
The effect of regularization:

Unlike the degree and the use of augmented features, regularized played a role, albeit small, in model selection. More specifically, it was observed that when using a λ that ranged in evenly spaced intervals from 0 to 0.9, 0.9 was always selected except for the use cases of China and Taiwan. It is very probable that a high λ was always selected in order to introduce bias and reduce the already high variance of the polynomial models, which is in large part owed to the size of the dataset. In those two corner cases, a λ of 0 was selected and it should be noted that those two countries had the most complex curves out of all the ones examined. As such, it seems preferable to reduce the regularization parameter than to utilize augmented features. Obviously, the fact that increasing parameters also increases AIC/BIC also plays a role, as reducing λ has no such effect.

The ranges of the BIC:

The dataset used for the fitting of models of all degrees per country was kept stable by utilizing a seed. As such, the BIC is not influenced by the randomness of the train, test split and the models can be compared against one another. In all cases, the BIC was in the range [-6,+7.1], which seems to be proportional to the complexity of the infection curve itself. However, comparing the BIC across countries is not advisable, as the dataset does change and there is no relationship between them.

Some indicative plots:



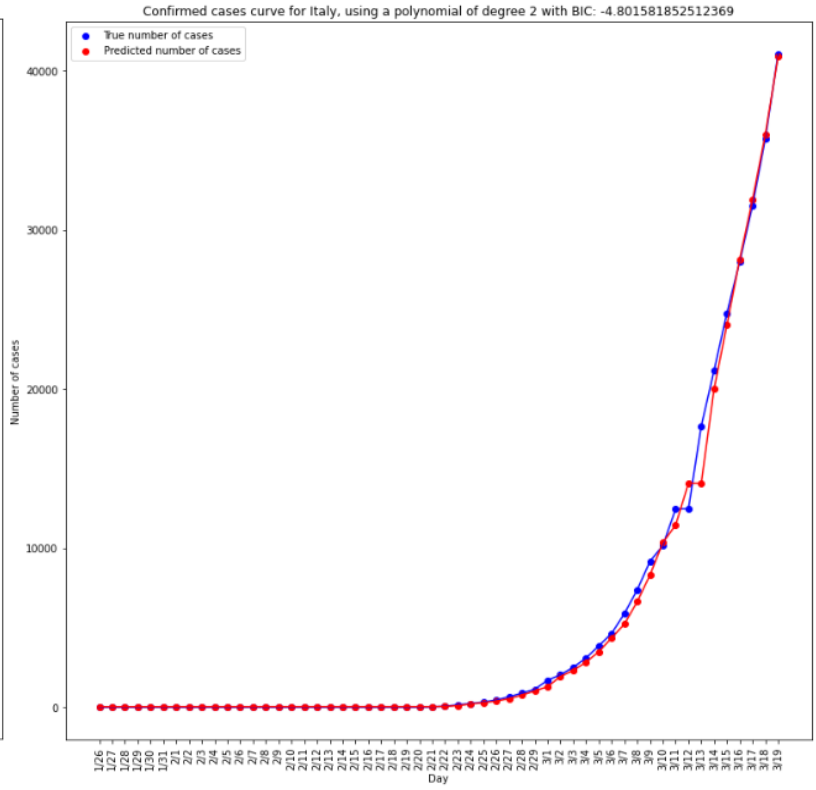
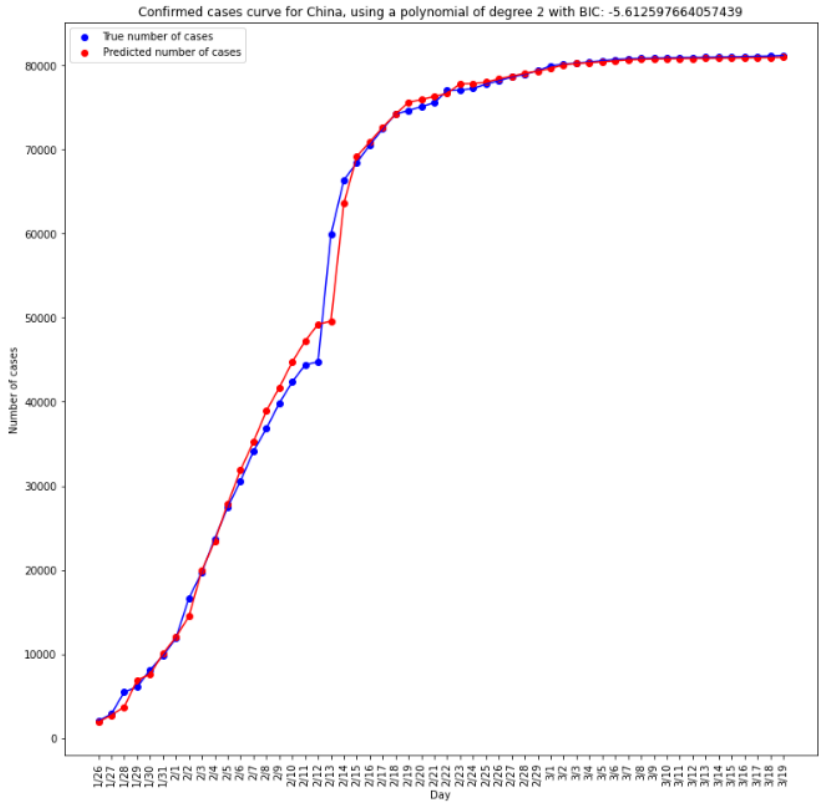


Figure 1: A model using the three previous days, the day index, as well as the confirmed deaths and recoveries for that day fitted to the data for Italy. While the optimal model selected is of degree 2, it contains many more features and as such is overfitting. This can be easily observed even visually, when the model has an abrupt oscillation from March 10 to March 16, whereas the actual infection curve is smoother. This is in stark contrast with the optimal model, which contains less redundant features and provides an overall better fit to the curve. The difference in the BIC is also quite pronounced, from 46 in the overfitted model to - 4.8 in the better fit one.



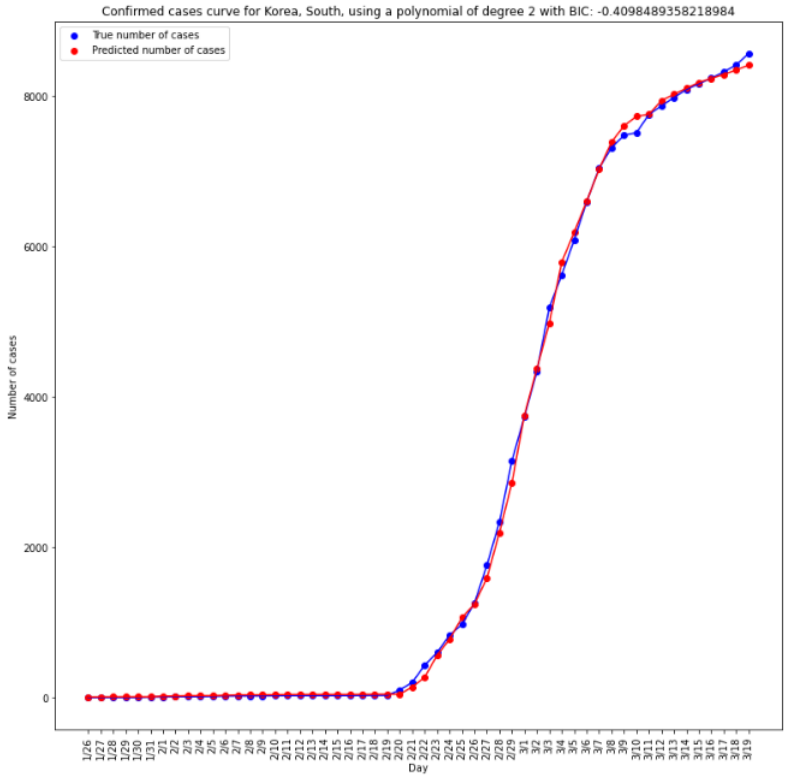
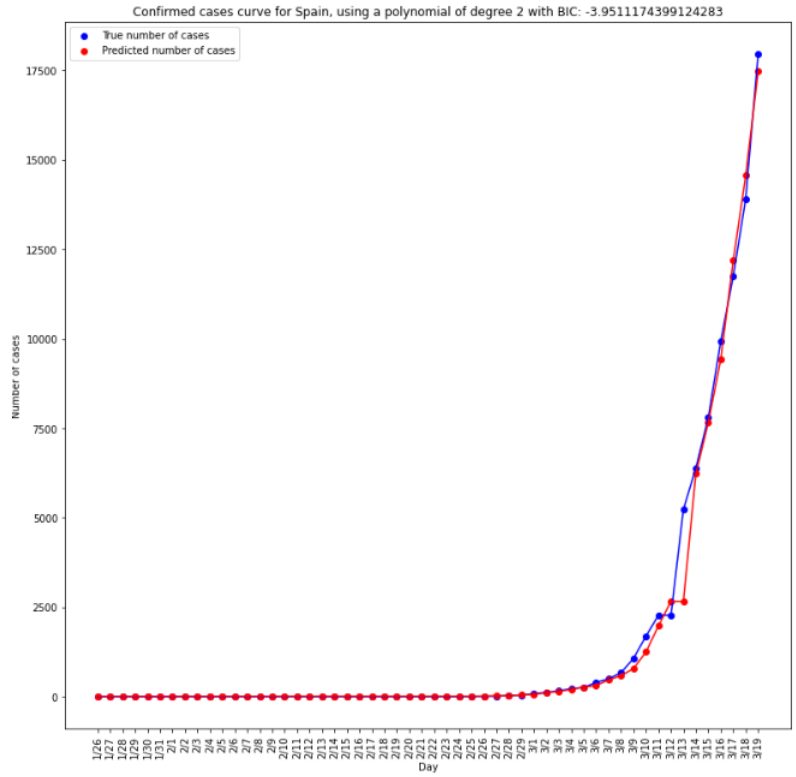


Figure 2: The cases of China and South Korea, where the curve has either flattened (China), or shows signs that it follows such a trajectory (South Korea). In these plots, only the optimal model is displayed, which provides a good fit for the data.



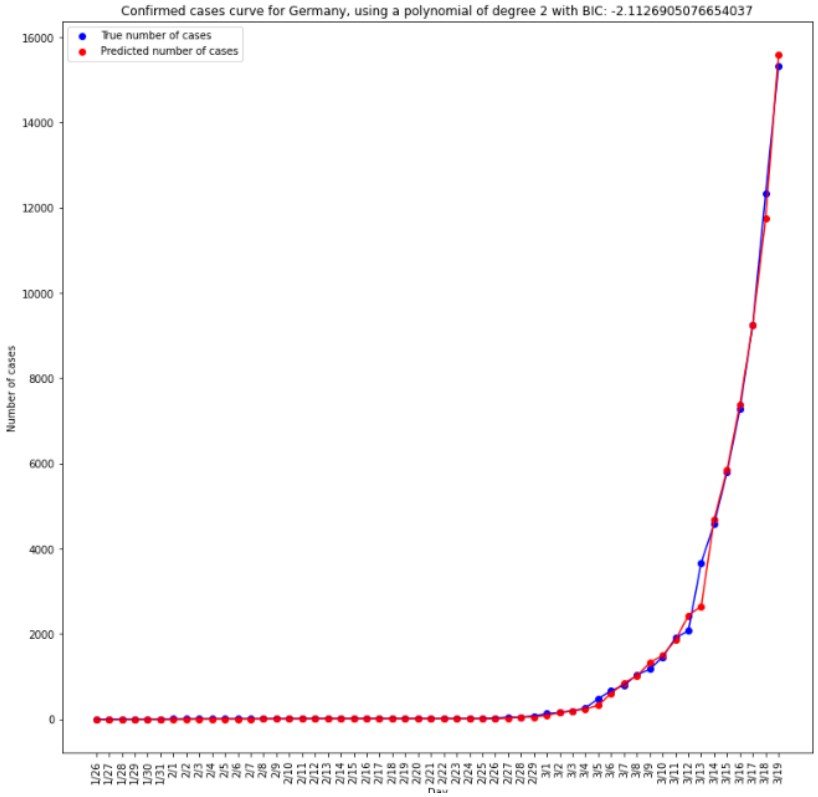


Figure 3: The cases of Spain and Germany, where the curve is still largely in its exponential phase and shows no signs of slowing. The curves are quite good fits for the data.

Discussion:

Seeing the curves, as well as the models, the pandemic does not seem to be slowing down its progression. Many countries are enforcing harsher and harsher measures and some seem unable to combat it. As such, a robust computational estimate for when the pandemic will subside is vitally important for medical staff, as well as industry and political leaders alike, since it will at least provide them with data and aid in planning for the future. Epidemiologists all around the world are tackling this problem constantly, but there seems to be no consensus on when the pandemic recede. This is also hampered by the fact that social distancing, as well as other measures, like travel bans, are not followed by the populace. Add to that the fact that the virus is still behaving in unpredictable ways and the reasons for which specialists have not been able to either combat the pandemic effectively or at least approximate a time when it may end become rather evident. Seeing as there supposedly will not be a vaccine until at least 2021, the chances that the quarantines and the travel bans will be extended seems extraordinarily high. Luckily, the virus' mortality is low, though this is only when not taking into account age and co-morbidities.

In this assignment, a small-scale attempt at modelling the dynamics of the epidemic was made. Needless to say, this work is quite limited. First, no extrapolation was performed for data outside date_x, except for Italy, though refactoring of the code in order to be able to do that more effectively seems trivial. Secondly, there are many features that may influence the epidemic, none of which are considered here. It would require expert domain knowledge in order to engineer plausible features that would aid in the model's predictive ability. As shown in the results section, simply creating polynomial features out of other provided features did not aid the models' predictive ability in any case. Furthermore, the lack of data hampers the use of more sophisticated algorithms for this specific problem. Many simplifications were made, such as concatenating all regions within a country into one nation-wide entry for confirmed cases, deaths and recoveries. Finally, there is a lot of work that can be done in the visualization department, as the plots produced in this assignment are quite crude. However, the code was written to be very extensible and further refactorization of imperative parts in functions will aid this endeavor further, possibly resulting in much higher-quality code down the line. The repository where the code will be maintained is here.

As for the model's extrapolation ability, which was not explored comprehensively, the model seems to overestimate the number of cases when extrapolating for the case of Italy. The figure is available in the notebook file that is attached with this report.

However, the outlook should not be so grim. When considering only the confirmed cases and assessing the mortality of the virus, it is quite likely overestimated. This is because most cases exhibit only mild symptoms that and can be handled by the body's own immune response. Only a fraction of the affected people needs hospital treatment and from them, only a fraction needs the Intensive Care Unit (ICU). In order to estimate the number of people affected in the general population, extrapolating from the number of confirmed cases is biased. First, the mortality of the virus needs to be estimated, by watching for the convergence of the ratios of deaths per closed cases and deaths per total cases. With some projection, one can make a rough estimate of this quantity. This estimate is quite biased on its own though. Knowing that 20% of the cases require hospitalization, 5% require ICU and about 2.5% require specialized machinery, the number of infected people in the population can be arrived at by working backwards from those estimates. It is more effective to estimate the number of infections from the general population, because it is not biased by the number of confirmed cases. However, a very important caveat is that it requires knowing general epidemiological statistics. As such, many estimates are based on the confirmed cases of the PCR results and take into account these biases when making estimates.

1. Bautista, E. *et al.* Clinical aspects of pandemic 2009 influenza a (H1N1) virus infection. *New England Journal of Medicine* (2010) doi:10.1056/NEJMra1000449.

2. Dawood, F. S. *et al.* Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: A modelling study. *Lancet Infect. Dis.* (2012) doi:10.1016/S1473-3099(12)70121-4.

3. Xie, M. & Chen, Q. Insight into 2019 novel coronavirus — an updated intrim review and lessons from SARS-CoV and MERS-CoV. *Int. J. Infect. Dis.* (2020) doi:10.1016/J.IJID.2020.03.071.

4. Wu, D., Wu, T., Liu, Q. & Yang, Z. The SARS-CoV-2 outbreak: what we know. *Int. J. Infect. Dis.* (2020) doi:10.1016/j.ijid.2020.03.004.

5. Hui, D. S. *et al.* The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health — The latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases* (2020) doi:10.1016/j.ijid.2020.01.009.

6. Wang, C. *et al.* The establishment of reference sequence for SARS-CoV-2 and variation analysis. *J. Med. Virol.* (2020) doi:10.1002/jmv.25762.

7. Liu, S., Zheng, Q. & Wang, Z. Potential covalent drugs targeting the main protease of the SARS-CoV-2 coronavirus. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa224.

8. Cagliani, R., Forni, D., Clerici, M. & Sironi, M. Computational inference of selection underlying the evolution of the novel coronavirus, SARS-CoV-2. *J. Virol.* (2020) doi:10.1128/JVI.00411-20.

9. Ortega, J. T., Serrano, M. L., Pujol, F. H. & Rangel, H. R. UNREVEALING SEQUENCE AND STRUCTURAL FEATURES OF NOVEL CORONAVIRUS USING IN SILICO APPROACHES: THE MAIN PROTEASE AS MOLECULAR TARGET. *EXCLI J.* (2020) doi:10.17179/excli2020-1189.

10. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* (2020) doi:10.1016/j.cell.2020.02.058.

11. Yan, R. *et al.* Structural basis for the recognition of the SARS-CoV-2 by full-length human ACE2. *Science* (2020) doi:10.1126/science.abb2762.

12. Prompetchara, E., Ketloy, C. & Palaga, T. Immune responses in COVID-19 and potential vaccines: Lessons learned from SARS and MERS epidemic. *Asian Pacific J. allergy Immunol.* (2020) doi:10.12932/AP-200220-0772.

13. Ahmed, S. F., Quadeer, A. A. & McKay, M. R. Preliminary identification of potential vaccine targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. *Viruses* (2020) doi:10.3390/v12030254.

14. Li, G. *et al.* Coronavirus infections and immune responses. *Journal of Medical Virology* (2020) doi:10.1002/jmv.25685.

15. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* (2020) doi:10.1038/s41591-020-0820-9.

16. Ge, X. Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* (2013) doi:10.1038/nature12711.

17. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* (2020) doi:10.1038/s41586-020-2012-7.

18. Menachery, V. D. *et al.* A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* (2015) doi:10.1038/nm.3985.

19. Zhang, C. *et al.* Protein structure and sequence re-analysis of 2019-nCoV genome refutes snakes as its intermediate host or the unique similarity between its spike protein insertions and HIV-1. *J. Proteome Res.* (2020) doi:10.1021/acs.jproteome.0c00129.

20. van Doremalen, N. *et al.* Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. *N. Engl. J. Med.*

(2020) doi:10.1056/NEJMc2004973.

21. Lewnard, J. A. & Lo, N. C. Comment Scientific and ethical basis for social-distancing interventions against COVID-19. *Lancet Infect. Dis.* (2020) doi:10.1016/S1473-3099(20)30190-0.

22. Bourouiba, L. Turbulent Gas Clouds and Respiratory Pathogen Emissions: Potential Implications for Reducing Transmission of COVID-19. *JAMA* (2020) doi:10.1001/jama.2020.4756.

23. Al-Tawfiq, J. A. Asymptomatic coronavirus infection: MERS-CoV and SARS-CoV-2 (COVID-19). *Travel Med. Infect. Dis.* (2020) doi:10.1016/j.tmaid.2020.101608.

24. Mizumoto, K., Kagaya, K., Zarebski, A. & Chowell, G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Eurosurveillance* (2020) doi:10.2807/1560-7917.es.2020.25.10.2000180.

25. Fung, S. Y., Yuen, K. S., Ye, Z. W., Chan, C. P. & Jin, D. Y. A tug-of-war between severe acute respiratory syndrome coronavirus 2 and host antiviral defence: lessons from other pathogenic viruses. *Emerging Microbes and Infections* (2020) doi:10.1080/22221751.2020.1736644.

26. Iacobucci, G. Sixty seconds on . . . anosmia. *BMJ* **368**, m1202 (2020).

27. Yang, X. *et al.* Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir. Med.* (2020) doi:10.1016/S2213-2600(20)30079-5.

28. Roussel, Y. *et al.* SARS-CoV-2: fear versus data. *Int. J. Antimicrob. Agents* (2020) doi:10.1016/j.ijantimicag.2020.105947.

29. Beberg, A. L., Ensign, D. L., Jayachandran, G., Khaliq, S. & Pande, V. S. Folding@home: Lessons from eight years of volunteer distributed computing. in *IPDPS 2009 - Proceedings of the 2009 IEEE International Parallel and Distributed Processing Symposium* (2009). doi:10.1109/IPDPS.2009.5160922.

30. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* (2020) doi:10.1016/S1473-3099(20)30120-1.