Tziporah Horowitz

Professor Adam Kapelner

Math 390.4

Friday 24th April, 2020

The Severity of COVID-19 can Only be Understood After its Defeat

1 Introduction

According to the World Health Organization, the term *pandemic* is defined as the worldwide spread of a new disease, regardless of its severity (*What is a pandemic?*, 2010). Although the world has reported pandemics as early as the second century, the modern world has seen an influx in global diseases due to increased population density and ease of travel (LePan, 2020). In recent years, the world has faced deadly outbreaks of influenza A (H1N1), SARS, Ebola, and Coronavirus. In early 2020, a new strain of Coronavirus, COVID-19 emerged in Wuhan, China quickly spreading to Europe and North America. However, the World Health Organization's reluctance to classify the virus as a pandemic led to insufficient precautionary measures taken by the global population. While the factors leading to a pandemic are understood, the severity of the COVID-19 is not. By using mathematical models similar that of pandemic diseases like influenzas and other Coronaviruses, the world can get a better grasp on what makes COVID-19 a dangerous pandemic.

Each year, strains of influenza are prevalent during "flu season" but seasonal flu is not classified as a pandemic because it does not occur simultaneously in both hemispheres.

Only when the virus appears "out of season" in a global capacity is it considered a pandemic (Kelly, 2011). Unlike pandemic influenza, viruses like SARS and MERS are closely related to COVID-19 by genetic structure. SARS, MERS, and COVID-19 are all part of a family of 7 RNA viruses called Coronaviruses that cause respiratory diseases in humans. Coronaviruses are usually mild, but these three are more severe than most strains (*Coronaviruses: SARS, MERS, and 2019-nCoV*, 2020). Pandemic influenzas usually have greater morbidity and mortality rates than Coronaviruses due to their high transmission rates. Nevertheless, COVID-19's reproduction number is uncharacteristically high and it in turn, poses a greater threat than SARS, MERS, and influenza A (H1N1). The purpose of this research is to develop a mathematical model that can explain what makes COVID-19 more dangerous than other pandemic outbreaks. By incorporating features used in previous models for pandemic risk, an accurate model can be made to extrapolate the severity of COVID-19.

2 Modeling the Severity of COVID-19

Human beings make assumptions and predictions based on prior experiences. Machines learn in a similar way by taking in input data and using it to produce outputs. Learning from data allows the machine to model natural phenomena that are not fully understood. George Box wrote, "Essentially, all models are wrong, but some are useful" (1987). Meaning, although models are merely approximations of reality and inherently contain error, they can be used to explain or predict a phenomenon. To be useful, a model needs metrics, ways to capture both the phenomenon and the features of reality causing it. Proper metrics are easily readable and unambiguous, have good resolution, and are

monotonic. Mathematical models use metrics to express ideas or abstractions in the form of mathematical equations.

Let the phenomenon $\boldsymbol{y}$ be the severity of COVID-19, determined by some true function $t(z_1, \ldots, z_t)$, where $t : z \to y$ and z_i are the true causal inputs. $t(z_1, \ldots, z_t)$ is not a model for the virus' severity because there is no way to determine the true function or its inputs. But, it is possible to find measurable features $x_1, \ldots, x_p$ that can approximate the information in $z_1, \ldots, z_t$. Although $z_1, \ldots, z_t$ are unknown, research suggests that the danger of any virus is determined by how infectious the disease is (z_1), how it affects infected people in the long term (z_2), and how equipped a locality is to fight the virus (z_3). Given proper metrics for features x_1, x_2, and x_3, there exists an ideal target function $f : x \to y$ such that the z_1, z_2, and z_3 can be estimated with some *error due to ignorance* (δ). Metrics for y, x_1, x_2, and x_3 will be discussed in the upcoming sections. While there is no analytical solution for finding the target function, f can be modeled through supervised learning.

Supervised learning is the process of using training data to identify an input-output relationship between the features and the phenomenon (Liu & Wu, 2012). Let $\mathbb{D} = \{\boldsymbol{X}, \boldsymbol{y}\}$ where $\boldsymbol{X} = [\mathbf{1}\ \boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3] \in \mathcal{X} = \mathbb{R}^{n \times p+1}$ and $\boldsymbol{y} \in \mathcal{Y} = \mathbb{R}^n$ be the training data and let $\mathcal{H}$ be the hypothesis set of candidate functions for f. There exists $h^* \in \mathcal{H}$ such that h^* is the closest possible model to f with some *misspecification error*. A learning algorithm, $\mathcal{A}$ uses $\mathbb{D}$ and $\mathcal{H}$ to produce a model $g(\boldsymbol{x}) \in \mathcal{H}$ near h^* with some *estimation error* (see eq. 1). The

$$\boldsymbol{y} = \overbrace{g(\boldsymbol{x})}^{model} + \underbrace{\overbrace{\big(h^*(\boldsymbol{x}) - g(\boldsymbol{x})\big)}^{estimation\ error} + \underbrace{\overbrace{\big(f(\boldsymbol{x}) - h^*(\boldsymbol{x})\big)}^{misspecification\ error} + \overbrace{\big(t(\boldsymbol{z}) - f(\boldsymbol{x})\big)}^{error\ due\ to\ ignorance\ (\delta)}}_{error\ (\epsilon)}}_{residual\ (e)} \tag{1}$$

goal of the model is to minimize the *residual error* (e), the difference between the true severity and the model's prediction. Residual error can be reduced by targeting the sources of error.

Mathematical models can only predict as long as the data remains stationary. Meaning, that the relationship between the phenomenon and the explanatory features does not change over time. This plays an important role in modeling the severity of COVID-19. Because new viruses mutate and their long term effects are unknown, severity is unpredictable. Therefore, the model can only be used for cross-sectional analysis and may need to be adjusted over time as new features of virus severity emerge.

3 Measuring the Severity of COVID-19

The term *severity* is typically defined as a subjective ordinal-categorical metric for how "bad" an event is. Because it is not a numerical measurement, it is difficult to quantify for analyses such as regression. However, by mapping severity to a function in $\mathbb{R}$, an objective composite metric can be derived. The World Health Organization defines pandemic risk $r(s)$ as the estimated annual probability of a pandemic having a greater severity than s standardized mortality units. The expected value of $r(s)$ can be used to quantify the pandemic's severity (Fan, Jamison, & Summers, 2018). Still, this method does not account for the economic loss caused by a pandemic and only estimates the expected loss of an uncertain event, not the actual loss of an occurring event. By adjusting the pandemic risk function given by the World Health Organization to follow a catastrophe modeling approach, the true severity of disease can be estimated.

Catastrophe models are used by actuaries to determine the financial impact of potential natural disasters. By using the estimated probability of loss, they can measure the direct, indirect, and residual losses of an event. Regardless of the type of catastrophe, these models always include metrics for predicting possible events, the intensity of an event, vulnerability to an event, and financial losses caused by an event (*Catastrophe Models*, 2020). While pandemic viruses are too complex to be modeled this way, catastrophe models can be used to evaluate the risk associated with one. Let $n = 134$ be the number of member countries and territories of the World Health Organization with more than 50 confirmed cases of COVID-19 as of April 4, 2020. $\boldsymbol{y} \in (0, +\infty)$ are the observed values of severity such that $y_i \in \mathcal{Y}$ is the per capita severity of the virus in a given location. The severity of the virus is defined by the World Health Organization's pandemic risk function adjusted for economic loss given a catastrophe model.

4 Determining the Features of Severity

A virus' risk of becoming pandemic is determined by both pathogen-specific factors and human population factors. The non-linear relationship between pathogenic features of independent viruses make it difficult to predict when a potential outbreak will occur or how severe it will be. However, it is possible to understand the severity of a specific pandemic by analyzing data that was collected during the outbreak. Data collected during an ongoing epidemic is often underreported and suffers from imprecise information, impacting predictions of outbreak size. It is therefore difficult to create a universal model for the severity of an infectious disease (Heesterbeek et al., 2015). Nevertheless, previous models for

specific viruses can be modified to make predictions about new pathogens.

4.1 Infectiousness (x_1)

Current models for infectiousness (z_1) focus on the basic reproduction number (R_0) of the virus. Let $\boldsymbol{x}_1 \in (0, +\infty)$ be the R_0 of COVID-19 in a given location. R_0 is a threshold parameter defined by the average number of cases transmitted to people by one individual. (Breban, Vardavas, & Blower, 2007). Because the transmission rate is dependent on social dynamics and interactions, R_0 can only be calculated under localized constraints (Cowling et al., 2010). Researchers have proposed various ways to compute the basic reproduction number, but all agree on the significance of of its calculation.

Clustering techniques are often used to determine the basic reproduction number of an epidemic. Following the MERS-CoV outbreak of 2012, Breban, Riou, and Fontanet proposed using a homogeneous branching process to create transmission trees for calculating R_0 via Bayesian analysis. This approach assumes that the number of secondary cases follow a Poisson distribution (2013). In the proposed model, basic reproduction number is determined using Breban et al.'s method for approximation. If R_0 is greater than 1, it is considered to have epidemic potential.

4.2 Morbidity Rate (x_2)

The long term physical effects of a pandemic (z_2) are most notably associated with the morbidity of the disease. The Center for Disease Control and Prevention uses metrics like the basic reproduction number to define morbidity (*Principles of Epidemiology*, 2012). Others however, define morbidity by its classical definition. According to Hernandez and

Kim, morbidity is the ratio at which an individual appears symptomatic or unhealthy (2020). Let $\boldsymbol{x}_2 \in (0, 1]$ be the average morbidity rate of COVID-19 in a given location. Common estimates for morbidity include disease prevalence and incidence; prevalence is the proportion of a population that have a specific attribute of a disease while incidence is the proportion of an initially uninfected population that develops the disease (*Principles of Epidemiology*, 2012). For the purpose of this paper, morbidity will be defined as the degree to which an individual experiences symptoms, expressed as a percentage. Morbidity differs from mortality in that it does not assume that the majority of people who contract the virus will die.

4.3 Hospital Preparedness ($\boldsymbol{x}_3$)

It is nearly impossible for a locality to determine which deaths were caused by a virus itself and which were coincidental (Jamison & Jamison, 2018). While the morbidity can be associated with fatality, the number of deaths caused by a disease can be exacerbated by insufficient hospital preparedness. In the context of a flu pandemic, Fineberg writes that the world is unprepared for global public-health emergency (2014). Hospital preparedness can account for part of the gap in disease severity between more developed and less developed nations. In more developed countries, most hospitals have plans for general emergency management; still, it is nearly impossible to prepare for a pandemic. Let $\boldsymbol{x}_3 \in [0, +\infty)$ denote the average hospital preparedness in a given location. A facility risk assessment can be used to evaluate the vulnerability of all areas in a hospital (*Emergency preparedness for healthcare*, 2019). Although there are many factors that determine hospital preparedness, this paper defines it as the number of respirators per hospital, given the nature of COVID-19.

5 Model Estimation

The purpose of this research is to develop a model that can explain what makes COVID-19 more dangerous than recent pandemics. While the true losses caused by the virus will not be known until the pandemic subsides, understanding factors such as basic reproduction number, morbidity rate, and hospital preparedness can help countries fight COVID-19. Using the metrics defined in previous sections, a cross-sectional analysis of COVID-19's current severity can be performed. Data can be obtained from the World Health Organization or government agencies like the CDC.

Provided dataset $\mathbb{D}$, the severity of COVID-19 can be approximated with a regression model using the *Ordinary Least Squares* algorithm (*OLS*). With the hypothesis set $\mathcal{H} = \left\{ \boldsymbol{w} \cdot \boldsymbol{x} : \boldsymbol{w} \in \mathbb{R}^{p+1} \right\}$ where p is the number of explanatory terms, a linear model $g_1(\boldsymbol{x})$ can be obtained in eq. 2 by minimizing the *sum of squared residual error* over all elements in

$$g_1(\boldsymbol{x}) = b_0 + b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + b_3\boldsymbol{x}_3 \tag{2}$$

$\mathcal{H}$. While eq. 2 explains the individual effects of the independent variables, it does not explain the interaction between insufficient hospital preparedness and morbidity. Pandemic viruses like COVID-19 are often made worse by underlying conditions, leading to an increased number in deaths attributable to the disease (*What is a pandemic?*, 2010). Moreover, those in less developed countries are more vulnerable to pandemics than those in more developed countries due to their lack of resources like medical capacity and proper hygiene (Jamison & Jamison, 2018). It is for this reason that a second model $g_2(\boldsymbol{x})$ was

developed in eq. 3 by expanding $\mathcal{H}$ to include interaction models

$$g_2(\boldsymbol{x}) = b_0 + b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + b_3\boldsymbol{x}_3 + b_4\boldsymbol{x}_2\boldsymbol{x}_3 \tag{3}$$

with a new design matrix:

$$\boldsymbol{X} = \begin{array}{c} \begin{matrix} \mathbf{1}_n & \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 & (\boldsymbol{x}_2 * \boldsymbol{x}_3) \end{matrix} \\ \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} & x_{21}x_{31} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{3n} & x_{2n}x_{3n} \end{bmatrix} \end{array}$$

When using *OLS*, the accuracy of the model can be measured by two metrics, R^2 and $RMSE$. $R^2 \in (-\infty, 1)$ is the proportion of $\boldsymbol{y}$'s variance that is explained by the model; an R^2 close to 1 implies that model predicts well, while a R^2 less than 0 indicates that the model performs worse than the null-model, $\hat{\boldsymbol{y}} = \bar{y}$. Unlike R^2, the *Root Mean Squared Error* ($RMSE$) has interpretable units and can be used to create a 95% predictive interval when e follows a normal distribution. Eq. 3 likely has a higher R^2 and smaller $RMSE$ than eq. 2 and therefore a closer fit, but increasing the complexity of the hypothesis set can lead to *overfitting*. Although overfitting produces low error by reducing in-sample noise, the model will not be precise when the data changes out-of-sample.

The problem of overfitting can be detected with *honest validation*, partitioning the data set $\mathbb{D}$ into training data $\mathbb{D}_{train} \in \mathbb{R}^{(n-k)\times(p+1)}$ and testing data $\mathbb{D}_{test} \in \mathbb{R}^{k\times(p+1)}$ (Abu-Mostafa, Magdon-Ismail, & Lin, 2012, p. 137-141). Honest validation can be performed to identify the optimal number of explanatory terms that will minimize the trade-off between

accuracy and overfit. Let $\mathbb{D}_{train}$ be a random sample of 80% of the observations in $\mathbb{D}$ such that $n_{train} = 107$ and let $\mathbb{D}_{test}$ be the remaining 27 observations. By fitting $\mathbb{D}_{train}$ to eq. 3, one can obtain the in-sample R^2 and $RMSE$. Applying the same OLS estimates to predict on $\mathbb{D}_{test}$ and obtaining the out-of-sample R^2 and $RMSE$ allows the modeler to observe the difference between the in-sample and out-of-sample standard errors of the residuals. It is then possible to gauge how well eq. 3 will perform with future observations. The same method can be applied to eq. 2 to determine which model has less *generalization error*.

It is important to note that honest validation does not guarantee that either of the models selected is the "best" model for selection. Rather, one can use a similar approach that splits the data into three parts: $\mathbb{D}_{train}$, $\mathbb{D}_{test}$, and $\mathbb{D}_{select}$, to perform cross validation on multiple models. By using a meta-algorithm for model selection one can perform k-fold cross validation on m models using $\mathbb{D}_{train}$ and $\mathbb{D}_{select}$ and obtain the out-of-sample $RMSE$. The model with the smallest variation in e, m_* can then be fit on $\mathbb{D}_{train} \cup \mathbb{D}_{select}$. Model selection can be applied to eq. 2 and eq. 3 to identify the better model of the two, $g_{final}(\boldsymbol{x})$. One can then perform validation to determine if $g_{final}(\boldsymbol{x})$ will overfit.

$g_{final}(\boldsymbol{x})$ will likely predict well when time is held constant; but due to the nature of viruses, it may fail in the future. The model will only be useful as long as it can predict the severity for new interpolated data. However, when new data is extrapolated, the model will likely fail because it is impossible to know if the relationship between severity and the features remains the same out of the range of the training data.

6 Conclusion

Real-world phenomena like a pandemic can only be explained by hypothesizing and observing data. Mathematical modeling allows one to do so by setting clear definitions and identifying associations. Although pandemic outbreaks cannot be predicted, observed viruses like COVID-19 can be modeled when time is held constant. Identifying the basic reproduction number, morbidity rate, and average hospital preparedness can help individuals better prepare themselves for the ongoing COVID-19 pandemic. Using interpolation, countries can predict on new data to determine the extent at which preventative policy should be implemented. While the true danger of COVID-19 is unknown, understanding the factors of its severity can help save thousands of lives.

References

Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H.-T. (2012). *Learning from data: a short course*. AMLBook.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. J. Wiley.

Breban, R., Riou, J., & Fontanet, A. (2013). Interhuman transmissibility of middle east respiratory syndrome coronavirus: estimation of pandemic risk. *The Lancet*, *382*(9893), 694 – 699. doi: 10.1016/s0140-6736(13)61492-0

Breban, R., Vardavas, R., & Blower, S. (2007). Theory versus data: How to calculate r0? *PLOS ONE*, *2*(3), 1 - 4. doi: 10.1371/journal.pone.0000282

Catastrophe models. (2020). National Association of Insurance Commissioners. Retrieved from `https://content.naic.org/cipr`$_t$*opics/topic*$_c$*atastrophe*$_m$*odels.htm*

Coronaviruses: Sars, mers, and 2019-ncov. (2020). Johns Hopkins Center for Health Security. Retrieved from `http://www.centerforhealthsecurity.org/resources/fact-sheets/pdfs/coronaviruses.pdf`

Cowling, B. J., Lau, M. S. Y., Ho, L.-M., Chuang, S.-K., Tsang, T., Liu, S.-H., ... Lau, E. H. Y. (2010). The effective reproduction number of pandemic influenza: "prospective estimation". *Epidemiology*, *21*(6), 842 - 846. Retrieved from `http://www.jstor.org/stable/20788237`

Emergency preparedness for healthcare. (2019). Premier. Retrieved from `https://www.premiersafetyinstitute.org/safety-topics-az/emergency-preparedness/emergency-preparedness/`

Fan, V. Y., Jamison, D. T., & Summers, L. H. (2018). Pandemic risk: how large are the expected losses? *Bulletin of the World Health Organization*, *96*(2), 129 – 134. doi: 10.2471/blt.17.199588

Fineberg, H. V. (2014). Pandemic preparedness and response — lessons from the h1n1 influenza of 2009. *New England Journal of Medicine*, *370*(14), 1335 – 1342. doi: 10.1056/nejmra1208802

Heesterbeek, H., Anderson, R. M., Andreasen, V., Bansal, S., De Angelis, D., Dye, C., ... (2015). Modeling infectious disease dynamics in the complex landscape of global health. *Science*, *347*(6227). doi: 10.1126/science.aaa4339

Hernandez, J. B. R., & Kim, P. Y. (2020). *Epidemiology morbidity and mortality.* StatPearls Publishing.

Jamison, D. T., & Jamison, D. T. (2018). *Disease control priorities. improving health and reducing poverty* (3rd ed.). World Bank Publications.

Kelly, H. (2011). The classical definition of a pandemic is not elusive. *Bull World Health Organ.*, *89*(7), 540 - 541. doi: 10.2471/BLT.11.088815

LePan, N. (2020). *Visualizing the history of pandemics.* Visual Capitalist. Retrieved from `https://www.visualcapitalist.com/history-of-pandemics-deadliest/`

Liu, Q., & Wu, Y. (2012). Supervised learning. In *Encyclopedia of the sciences of learning* (p. 3243 - 3245). Boston, MA: Springer US. doi: 10.1007/978-1-4419-1428-6_451

Principles of epidemiology. (2012). Centers for Disease Control and Prevention. Retrieved from `https://www.cdc.gov/csels/dsepd/ss1978/lesson3/index.html`

What is a pandemic? (2010). World Health Organization. Retrieved from https://www.who.int/csr/disease/swineflu/frequently_asked_questions/pandemic/en/