AM115 Workshop: COVID19 case study

Mathematical modeling has been used in a variety of ways in the fight against the COVID-19 pandemic. In particular, to provide information for devising public policy and planning responses, projections of the pandemic trajectory are made by many models[1]. Many use extensions of the SIR model, some use purely data-driven machine learning approaches, and some (fewer) use individual agent type of approaches (mostly for smaller communities). The SIR-type of models tend to do better in longer-term projections and some modelers combine them with purely machine learning approaches for their projections.

In this workshop, we shall take a closer look at a particular model by the Institute for Health Metrics and Evaluation (IHME) at University of Washington. The IHME model is described in this paper "Modeling COVID-19 scenarios for the United States", Nature Medicine, 2020, accessible at https://www.nature.com/articles/s41591-020-1132-9, which is the source for the material (including the data) used in this workshop. Most of the details are in the supplemental information.

The IHME model, roughly speaking, has the following components:
1. Use existing data and literature to constrain the parameters of an SIR-type model. In particular, the rate of spreading is estimated as a function of time.
2. The time-dependent rate of spreading is fit to covariates such as mobility and mask use.
3. The covariates are fit to policy mandates.
4. The SIR-type model is then used to map out the trajectories of the pandemic for different scenarios of public policy.

The IHME model extends the basic SIR model and includes five compartments: susceptible (S), exposed (E), infected and pre-symptomatic (I_1), infected and symptomatic (I_2), recovered/removed (R). The sum of all the compartments is the total population N. The movements from one compartment to another is described by Eq. (1) below. The parameters are named consistently with the IHME paper instead of with what we had in our SIR workshop.

$$\begin{aligned}
\frac{dS}{dt} &= -\beta(t)\frac{S(I_1+I_2)^\alpha}{N} \\
\frac{dE}{dt} &= \beta(t)\frac{S(I_1+I_2)^\alpha}{N} - \sigma E \\
\frac{dI_1}{dt} &= \sigma E - \gamma_1 I_1 \\
\frac{dI_2}{dt} &= \gamma_1 I_1 - \gamma_2 I_2 \\
\frac{dR}{dt} &= \gamma_2 I_2
\end{aligned} \tag{1}$$

[1] For some examples of the models, see this website by the Center for Disease Control (CDC) https://www.cdc.gov/coronavirus/2019-ncov/covid-data/mathematical-modeling.html

The parameter $\alpha \leq 1$ adjusts for imperfect mixing within the population; clustering reduces the efficiency of the disease spread. The other parameters are rates for moving from one compartment to another. The range that IHME gives for the parameters based on the literature is $\alpha \in [0.9,\ 1]$, $\sigma \in [0.2,\ 1/3]/\,day$, $\gamma_1 = 0.5/\,day$, $\gamma_2 \in [1/3,1]/\,day$. In the IHME implementation, these parameters are randomly drawn from their uncertainty range and then kept constant in time. A large number of draws are done to characterize the uncertainty. In this workshop, for simplicity, we will take α=0.95/day, σ=0.25/day, γ_1=0.5/day, γ_2=2/3/day. But you are welcome to try random draws from the distributions if time permits.

While things like the recovery rates may depend on the care the patients receive, which may change over time, it is reasonable to approximate the above rates as constant in time and focus on the change in β, the time dependence of which is explicitly acknowledged in Eq. (1).

To estimate β(t), we need estimates of the numbers of new infections. The numbers of new cases each day reported by health departments however are not reliable estimates due to inadequate testing, particularly in the early days of the pandemic, as we can all recall. What the IHME group did was to use the daily death numbers, which is more reliable, to estimate the infection numbers. In a number of communities around the country, seroprevalence studies were carried out, where people are randomly tested to determine the prevalence of the disease for each age group. This information is then combined with the number of deaths for each age group in these communities to determine the age-specific Infection-Fatality Ratio (IFR). The IFRs are assumed to be the same in all states and used together with the daily death numbers to estimate the number of new infections, with adjustments for the time lag from infection to death. Their estimates for each state is included in daily_infections.csv[2]. Note that at the time of their research on which this paper was based, estimates based on actual data were only available before Sep. 21, 2020. The numbers after that date are model projections based on their reference scenario. You can also read out the population of each state from US_states_population.csv. An example code to read the data is in covid_code4student.m and covid_code4student.ipynb.

Activity 1: using the number of new infections given, determine β(t) using Eq. (1)

The estimated daily infections is f(t) times one day, where we define

$$f(t) \equiv \beta(t) \frac{S(I_1 + I_2)^{\alpha}}{N} \tag{2}$$

1a. Given f(t), integrate Eq. (1) using an ODE solver to find S(t), I_1(t), I_2(t). We shall assume f is constant within each day.

1b. Interpolate the solution to daily resolution, then use Eq. (2) to solve for β(t). My solution of β(t) is included in betas.csv for you to compare with. I started the simulation from Feb. 1, 2020.

[2] If you want to compare the estimated daily new infections with the reported new cases, take a look at the file IHME_CaseDeathHospitalization.csv. Remember the relevant column is no longer named "mean". This comparison is not part of the workshop, do it only if you are curious and after you are done with the other parts.

The next step is to determine how covariates such as mobility and mask wearing affect the spread rate β(t). Mobility was estimated by IHME based on mobile phone user information collected by companies such as Google, Facebook and others and is given in mobility.csv. Mask use is estimated based on survey data asking people how often they wear masks when outside their homes, and given in mask_use.csv. The IHME paper also used other covariates such as testing and seasonal pneumonia, but we will limit ourselves to mobility and mask use, as these two are the most important factors.

Activity 2: estimate the effect of mobility and mask use on the spread rate $\boldsymbol{\beta}$(t)

2a. Perform a linear regression of log β(t) against mobility and mask use. Because different states have different baseline values of β, depending on the demography, population density, prevalence of smoking, air pollution, and altitude, etc., it is best to use a mixed effect model. Here, for simplicity, we will remove the mean of each variable for each state before pooling data of all states together for the linear regression. We shall also exclude the early days of the pandemic and start from say March 11, 2020, given that the earlier infections could be affected more by imported cases instead of community spread[3]. Even with this exclusion, states with very small case numbers initially can still have unrealistic values of β. So look out for them.

2b. What assumption are we making by regressing log β(t) instead of β(t)?

2c. Examine the fit and the contributions from mobility and mask use. Write down your thoughts about the fit. If time permits, examine whether the assumptions for Ordinary Least Squares are valid.

Lastly, for projections of the pandemic trajectory, we would need to determine how policies affect the covariates such as mobility, with which, we would then explore the consequences of different policies (such as reopening schools when the infection rate is less than x%). While we don’t have time to fit these in the workshop, if you are interested, take a closer look at the paper to get a sense of how mathematical modeling is used in the real world and the uncertainties associated with it.

[3] For some context on the timeline, President Bacow emailed the Harvard community on March 10, 2020 for the dorms to be vacated by March 15 and Massachusetts’s stay-at-home order was issued on March 15, and started on March 17.