CSE 8803 EPI: Data Science for Epidemiology, Fall 2022

Lecturer: B. Aditya Prakash
Scribe: Shangqing Xu, Akshath Shvetang Anna

August 30, 2022
Lecture #3 : Modeling

---

1 Summary of Last Lecture

Last lecture, we learned that the field of epidemiology studies the spread and control of diseases, viruses, concepts, and more throughout populations or systems. We then looked at the history of epidemiology and discussed some notable examples of computational epidemiology such as Bernoulli and Jenner's study of the value of variolation during the Smallpox epidemic and Jon Snow's spatial analysis of cholera data to identify the source of contamination. We considered the modern risks of pandemics in the context of the Covid-19 pandemic. We then took a look at data science and how it can help us answer questions as epidemiologists and the challenges associated. Today, we will look at one tool from this space - modeling.

2 A two paragraph summary of this lecture

In epidemiology, we create models to forecast what we think will happen with an infectious disease in the short-term and long-term given data that is sometimes limited and noisy. We ask a series of questions about the purpose, underlying assumptions, uncertainty, and generalizability of these models and what data is being used in order to create better and more useful models. These models come with properties that conflict and trade off with one another including accuracy, transparency, and flexibilty. One model of note is Bernoulli's Smallpox Model which describes the dynamics of a population that contains members who are susceptible and immune to a disease and how this changes as portions of that population are immunized through variolation. This model shows how measures such as immunization can lead to drastically larger life expectancies.

Another important model is SIR and the related family of ODE models that model the dynamics of a population that contains members who are susceptible to, infected with, and recovered from an infectious disease (as well as exposed but not infectious in the case of SEIR). An implicit solution of these models reveals an important term called R_0 whose value allows us to understand whether an epidemic will die out or become large. Many decisions can be made to change the value of R_0.Overall, ODE models are considered to be the workhorse of epidemiology as they are empirically-backed, easy to extend, and good for long-term forecasting. However, some of their limitations lie in their reliance on simplifying assumptions that may not reflect reality.

3 Why Models?

"Essentially, all models are wrong, but some are useful" - George E. P. Box

This quote emphasizes that we should be careful when assessing the outcomes of models. Because models are greatly influenced by the assumptions we make, they almost never

match up *exactly* with reality and should not be taken as gospel. However, by carefully constructing models with good assumptions, we can still learn a lot from them, whether qualitative or quantitative.

3.1 Limited and Noisy Data

We don't know how many people are actually infected or not at any given time, and there is always only have fraction of people who are effected that we can surveil. Thus, we sometimes need to rely on approximations and simplifying assumptions.

3.2 Forecasting and What-If Scenarios

Models can either be used for predicting outcomes (forecasting) or more generally to gain an understanding of a phenomena. We can achieve the latter through testing different What-If scenarios. For instance, what happens if certain measures are implemented, and how will these effect transmission? We need a causal model to figure this out. In general, we need a model of how the world works to answer the questions we have. This can help us guide our decision-making.

3.3 Providing guidance and intervention possibilities

As there are many real-world factors that need to be considered while doing predictions, we need an editable model so that we can guide or explicitly interfere the inner mechanisms how model performs onto real-world scenarios.

4 5 Questions to Ask about Model Results

4.1 What is the purpose and time frame of the model?

Some questions have models that need short-term forecasting and some need models with long-term forecasting. That's why it's important to ask why you are creating this model. And what is the ultimate goal?

4.2 What are the basic model assumptions?

This requires some work to understand things like what are the parameters, contact patterns, etc. that allow the model to exist as is without having to consider more complexities.

4.3 How is uncertainty being displayed?

This is very important because data is noisy, you have assumptions, and the situation is volatile. Without an understanding of how much uncertainty there is, it's hard to interpret the results.

You have to ask questions like whether confidence intervals are calibrated properly and what parameters are being varied. These are tough asks, and lots of models don't get them answered fully.

4.4 If the model is fitted to data, what data is being used?

How do we understand which data points are most useful? We need to understand the context of the model and ensure our data is appropriate.

4.5 Is the model generalizable or does it just reflect a particular context?

A model might be built for Atlanta, and might be less useful somewhere else. We need to understand the underlying assumptions made for things like age, demographics, mobility patterns, etc. and see if we can generalize the model to other situations accordingly.

5 Conflicting Properties

5.1 Accuracy

If a model is too detailed, it is hard to generalize it. Accuracy is important in general, but what metrics are most important (precision, recall, etc.)? We also have to consider the differences between short-term and long-term accuracy.

5.2 Transparency

Can we understand why the model is making certain predictions or decisions? This also suffers if you make the model too detailed or have too many parameters. It is important to ensure the model does not become a black box with no way of knowing why it makes the predictions that it makes.

5.3 Flexibility

If a model is too simple, it won't be flexible. This is to say that if a model has too little parameters, it can't produce meaningful results, especially when there is a variance in the situation. On the other hand, if a model is too complex and over-fitted to a particular situation, it'll also lack the flexibility of being generalizable.

5.4 Overall

These are properties we will focus on as we go over generative, theory-based models. We will talk about ODE-based, Agent-based, Multiscale/Metapopulation-based, and Network-based models. These are called *mechanistic models* and they are straightforward but not the most accurate. Later we will also discuss statistical models, which are more in vogue at the moment.

6 Bernoulli's Smallpox Model

6.1 Context

The idea of variolation was well known at the time. Some people have mild cases, so we take samples from them and use them to inoculate target patients to give them mild cases as well as immunity. This was common in Asia, England, US, but not France. Bernoulli wanted to compare the benefit of variolation vs immediate risk of dying. How do you quantify this?

6.2 Setup

The population is divided into

- Susceptible (not yet infected)
- Immune (immunized for life after one infection)

Assume those infected with smallpox either

- Die instantaneously with probability a
- Get life immunity

This is a rough modeling of what happens in real life.

6.3 Model

$$I(t) = x(t) + z(t) \tag{1}$$

- $I(t)$ = probability of survival till age t
- $x(t)$ = prob of never getting infected
- $z(t)$ = prob of immunity

Note that $x(t)$ is hard to determine because contact patterns are different. This is more true for homogenous populations, so that will be part of the simplifying assumptions. Assume the probability of anyone in $x(t)$ getting smallpox is a constant b. We have the following:

$$\frac{dx}{dt} = -bx(t) \tag{2}$$

$$\frac{dw}{dt} = (1-a)bx(t) \tag{3}$$

$$x(t) = \frac{w(t)}{(1-a)e^{bt} + a} \tag{4}$$

a and b can be roughly inferred from observational data

6.4 Results

If all children were variolated at birth

- Population will be 14% larger at age 26
- Life expectancy increases from 26.58 to 29.75

But this doesn't include the risk of variolation itself. Assume 2% die from variolation. This reduces life expectancy gain by 1 month, so still net almost 3 year increase in life expectancy. We can extend analysis to secondary spread infections from variolation, but this doesn't really change the bottom line that you can get almost 3 years additional life expectancy with variolation from birth. We have to make these sorts of complex tradeoff judgements, and models like this help us do that.

Note that there are lots of simplifying assumptions that ignore demographics or conditions of people, but this is still useful for providing general guidance.

In figure 1, we can vizualize life expectancy increase due to immunization:

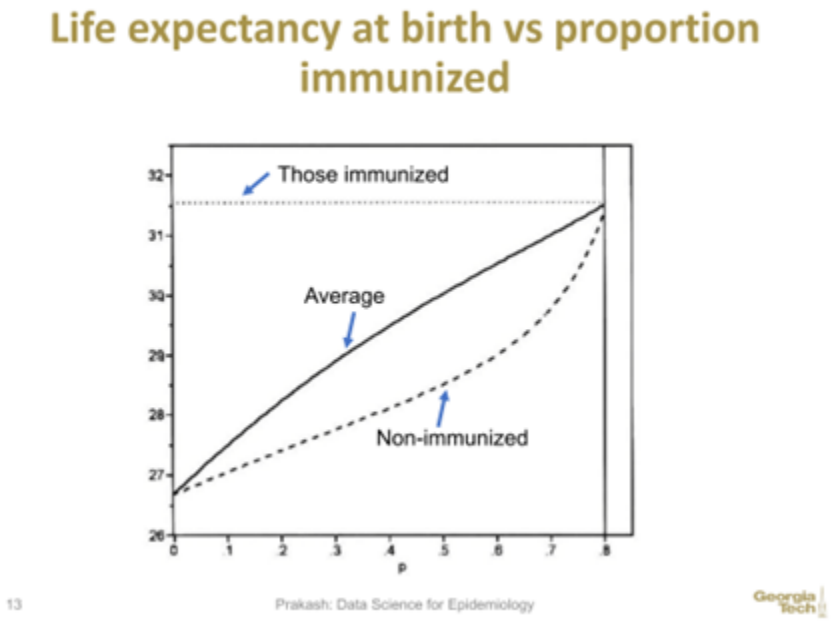

Figure 1: The above shows how average life expectancies for immunized and non-immunized portions of the population increase as a larger proportion gets immunized. Notice how at some point (80% immunized) the non-immunized portion has the same life expectancy as the immunized portion. This shows when herd immunity has been achieved. Thus the more people that get vaccinated, the greater the chances are for everyone to be safe.

7 ODE Models: SIR

7.1 Setup

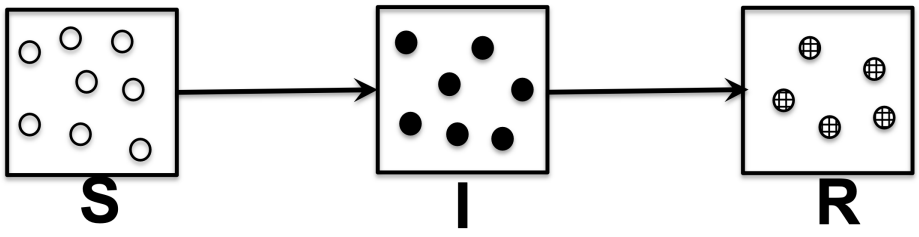

Figure 2: An abstract visualization of SIR models. Healthy people (S group) can turns into infected (I group) and finally get recovered (R group).

As shown in Figure 2, SIR is one of the most simple models that categorizes people into three categories:

- Susceptible: healthy, but can get infected
- Infected: can infect others through contact
- Recovered: can't infect others (dead or no longer at risk of being infectious)

This model base on three assumptions:

- Perfect mixing: Any infected person can infect any susceptible person (meaning all people come into contact with each other)
- The total population remains constant across time, meaning no birth or deaths.
- The properties inside the model are deterministic (instead of a probabilistic distribution).

7.2 Model

$$\frac{dS}{dt} = -\beta SI \tag{5}$$

- The above represents number of susceptible who are getting infected
- The change in number of susceptible people is inverse to the change in number of people getting infected as expected

$$\frac{dI}{dt} = \beta SI - \delta I \tag{6}$$

- BSI is number of new infections
- δI is number of infected that are cured or removed
- SI is number of times susceptible and infectious meet

The above equation represents the rate of change of infectious people. Notice that if there are I infected people, they are cured or getting removed at rate δ, so we have:

$$\frac{dR}{dt} = \delta I \tag{7}$$

7.3 Force of Infection

$$F = \lambda S \tag{8}$$

where F is the Force of Infection and λ is number of infected people.

Kinds of transmission:

- Mass-action transmission $\lambda = \beta I$
- Density-dependent $\lambda = \beta I/N$
  - As density increases so does transmission
  - Depends on density of infected people rather than total number of infected people
  - Consider $\beta' = \beta/N$ to update SIR

7.4 Solving SIR

There is no closed form solution, but numerical integration produces something like figure 3
As an example of applying SIR to real cases, see figure 4

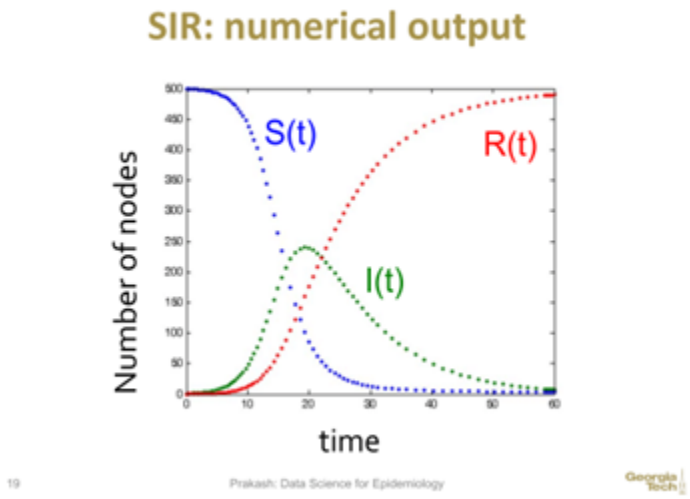

Figure 3: As seen above, SIR doesn't naturally model multiple waves of a pandemic without external pushing. View a more interactive example at the following link: http://epirecip.es/epicookbook/chapters/sir/python

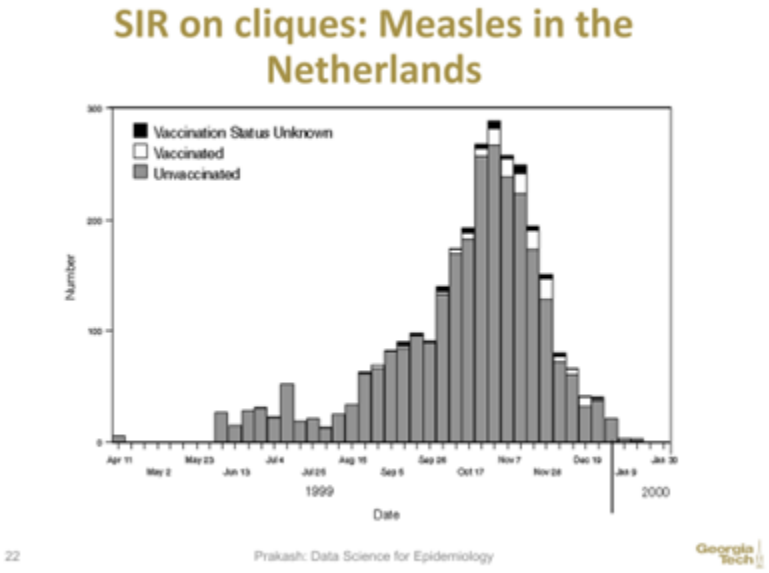

Figure 4: The goal is to find β and δ where I fits the curve above

7.5 SIS Model

Assume perfect mixing

$$\frac{dS}{dt} = -\beta SI + \delta I \tag{9}$$

$$\frac{dI}{dt} = \beta SI - \delta I \tag{10}$$

This results in an endemic state where there is an equilibrium with infected still existing. In SIR, infected eventually drops to 0, making SIS more useful for endemic states and SIR more useful for modeling diseases that burn through the population. Possible extensions of these models include:

- Birth/death rates
- Variable contact rates
- Age-structure models
- Make things stochastic
- Multiple viruses/diseases

7.6 SIR with Birth and Death

$$\mu : birth/deathrate \tag{11}$$

$$\frac{dS}{dt} = -BIS + \mu\left(I + R\right) \tag{12}$$

$$\frac{dI}{dt} = BIS - \gamma I - \mu I \tag{13}$$

$$\frac{dR}{dt} = \gamma I - \mu R \tag{14}$$

7.7 General Scheme of SIR

See figure 5 for an overview of the general scheme of SIR extensions

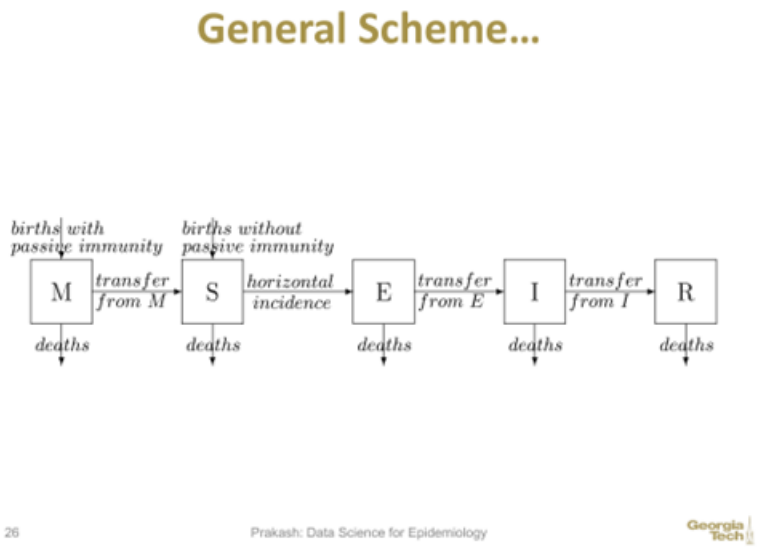

Figure 5: Nonlinearity introduces complexity and more interesting model behaviors

7.8 SEIR: SIR with Exposed State

People move from S to E to I. E is exposed state (Exposed to infections but not showing symptoms/infectious yet). This is 1-4 days for Flu and 2-14 for COVID.

$$\frac{dS}{dt} = -\beta IS \tag{15}$$

$$\frac{dE}{dt} = \beta IS - \sigma E \tag{16}$$

$$\frac{dI}{dt} = \sigma E - \gamma I \tag{17}$$

$$\frac{dR}{dt} = \gamma I \tag{18}$$

You can add even more detail by baking complexities into the constants.

7.9 SIR: Implicit Solution and R_0

$$S(t) = S(0)\, e^{-R_0(R(t)-R(0))} \tag{19}$$

$$R_{\text{inf}} = 1 - S(0)\, e^{-R_0(R_{\text{inf}}-R(0))} \tag{20}$$

$$R_0 = N\beta/\delta \tag{21}$$

R_0 is the reproductive number. There is a threshold phenomenon with this constant. Observe that

$$\frac{dI}{dt} = I(\beta S - \delta) \tag{22}$$

This implies that:

- The rate of change of infected is less than 0 if $S(0) < \frac{\delta}{\beta}$
- This means if $S(0) < \frac{\delta}{\beta} = \frac{1}{R_0}$, the epidemic will die out
- We will have a large epidemic if and only if $R_0 > 1$

Because ODEs are network based models where everyone is assumed to be in contact with everyone else, it is especially important to pay attention to and estimate R_0. Many decisions can be made based on changing the value of R_0. Immunization, for example, seeks to reduce $S(0)$ to below $\frac{1}{R_0}$.

Figure 6 shows how disease dynamics change as R_0 changes. We see as R_0 increases, epidemic last longer but the peak infected population would be greatly decreased.

Note that R_0 is not easy to estimate and may be miscalculated. SARS, for example, was estimated in hospitals and assumed perfect mixing which caused R_0 to be estimated to be much higher than reality suggested.

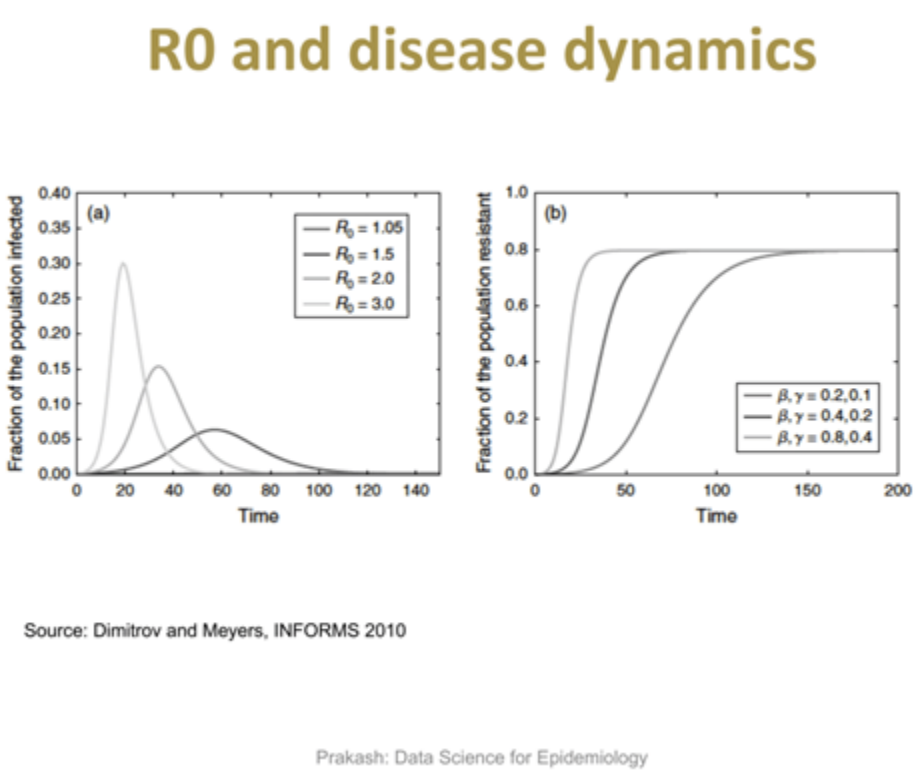

Figure 6: Simulations of infected percentage as R_0 and other hyperparameters change.

7.10 Pros and Cons of ODE models

ODE models are considered to be the workhorse of epidemiology because

- There have been many success stories over 100 years
- They are easy to extend and build upon
- We have good numerical solvers
- They sometimes can be solved analytically
- Are good for long term forecasting

However, a few of their limitations include

- There is a lack of heterogeneity. Within in each class, all the "people" are treated uniformly.
- Have too many simplifying assumptions to perform short-term forecasting
- Don't properly quantify interactions between people
- Are difficult to refine
- Don't account for how humans adapt and change their behavior

- Don't provide the information neeeded to properly design implementable interventions

Take sars for example. It has really high R0 (2.2-3.6) and spread across different countries, but finally creates few infections. One possible reason is that SARS infections are estimated in hospitals. In real-world scenarios, full mixture and corresponding infections are not easy to achieve.

References