<1>Chapter 5: Modeling infectious disease epidemics

In previous Chapters, we ignored a critical aspect of modeling some major diseases: the infectious nature of many diseases. For infectious diseases, the risk of getting the disease is related to how many people are infectious at a given time; the more infectious people in the area, the higher the risk of infection among susceptible people. In a typical Markov model, we can't account for this basic feature of infectious diseases, because the risk of moving from one state (healthy) to another state (diseased) is assumed to be constant. In this Chapter, we introduce a simulation modeling framework that has been used for decades to simulate infectious disease epidemics.

<2>Using hazard functions

Let's revisit an example we first started in Chapter 4: the case of a businessperson who is selling marshmallow candies ("peeps") and wants to know how many candies to order from the factory to optimize her profit.

A savvy businessperson could do a simple survey at the peep warehouse by asking the warehouse manager, "How long does the average peep stay on the shelf after it arrives, before it gets ordered and shipped off to the grocery store?" If the answer is 3 days, then what is the average daily per peep rate of leaving the warehouse? Well, it's 1/3 per day. (That is, 1 peep/3 days = one-third of a peep leaving per day). Conversely, if you have a really big fan club of customers, and peeps fly off your shelf at the rate of 5 per day on average, then what's the average *survival time* of one peep in your warehouse? It's 1/5 of a day (1 day/5 peeps = one-fifth of a day on the shelf per peep).

Therefore, the average survival time of peeps in your warehouse is:

$$E(T) = \frac{1}{\mu}$$

[Equation 9.1] ,

where E means the "expectation" (or average) of the time T, which we define as the amount of time a peep survives on the shelf in your warehouse, and μ is the average per peep rate of leaving the warehouse.

Let's find a function to describe the probability that a peep stays on the shelf for at least t units of time. We can write that as $\Pr\{T>t\}$. This is often called a *survival function*, designated *S(t)*. What could this function be? As we first derived in Chapter 4, we can reason that the average shelf life in the warehouse, *E(T)*, is just a sum of 1day*Pr{peep on shelf 1st day}+1day*Pr{peep still on shelf 2nd day}+1day*Pr{peep still on shelf 3rd day}+…all the way to infinity days (by which time, presumably, the probability of survival has reached 0). In other words,

$$E(T) = \int_0^\infty \Pr\{T > t\}dt = \int_0^\infty S(t)dt = \frac{1}{\mu}$$

[Equation 9.2]

We further reasoned in Chapter 4 that a function *S(t)* that can satisfy this equation, such that its sum from 0 to infinity is $1/\mu$, is the exponential function:

[Equation 9.3]

Hence, we have a function to describe the probability that a peep survives to time t, shown in Equation 9.4:

$$S(t) = e^{-\mu t}$$

[Equation 9.4]

To operationalize this equation, we can reason that if the typical duration of shelf-life for a peep is 3 days, then the probability a peep survives 7 days is $e^{-(1/3)*7} = 0.097$. because 1/3 is the rate of leaving the shelf (1 peep/3days) and t is 7 days. Hence, there is a 9.7% chance of a peep surviving a week on the warehouse shelf.

Conversely, we can use Equation 9.4 to convert from a metric of time to a probability of

survival. For example, suppose we want to know how much time it would take a peep's survival probability to be just 25%. We have to solve for t when $S(t) = 0.25$. So $0.25=e^{-(1/3)*t}$, and taking the natural logarithm of both sides, we get $ln(0.25) = -(1/3)*t$, which gives $t = 4.2$ days.

How does optimizing our warehouse of peeps help us to find solutions to stop infectious disease epidemics? We first have to determine how to adjust our equations to address the case of *fluctuating demand*. In other words, if we will simulate an infectious disease where the number of people infected per unit time will increase when more people are infectious, we no longer have a constant rate λ but rather have fluctuating rate of infection that could vary widely across time periods. Hence, the survival function—the probability of surviving at least to time t—will not simply be an exponential function of the constant λ summed over all times up to t (i.e., λt), but rather the exponential function of individual rates for each time, $\lambda(t)$, summed over all times up to time t:

[Equation 9.5] $$S(t) = e^{-\int_0^t \mu(\tau)\,d\tau}$$

The function $\lambda(t)$ is commonly referred to as a *hazard function*, meaning the hazardous probability that a peep will "end its life" at the warehouse (by leaving the warehouse, not via peep suicide) at a given time t, given that it survived in the warehouse until time t.

Now that we have a survival function that can accommodate a fluctuating rate across time periods, we can convert our model of peep consumption into a model of infectious disease epidemics.

<2>The Kermack-McKendrik Model

Suppose we want to conceptualize the process of an epidemic using the classic *Kermack-McKendrick model*, which is also known as the *SIR model* of disease because it has three states that people (a.k.a. "peeps") can fit into (Figure 5.1). People get shipped around between these states when they are infected with the disease being modeled (moving them from the warehouse of susceptible people to the warehouse of infected people), or recover from infection (moving from the warehouse of infected people to the warehouse of recovered people). The model assumes that the infection is not fatal, that infected people are infectious people, and that recovered people have lifelong immunity. We'll modify the assumptions later in more complex models, after we learn the basics. (Note that in some infectious disease literature, states are referred to as *compartments*, and the SIR model as a *compartmental model*).

[INSERT FIGURE 5.1 HERE]

How do we use this simple structure to answer complex questions about controlling infectious disease epidemics? One systematic approach is to label and define all of the rates of flow between states of the model, so that we can construct a series of equations to describe the process of infection and recovery.

Based on Equation 9.1, we can reason that the rate of infection can be expressed as the reciprocal of the expected time until a person moves from the susceptible state to the infected state (let's call that $E(T_S)$). The expected time until a person moves from susceptible to infected is the average length of time between birth (which we assume is when a person enters the susceptible state) and the time a person gets infected. Similarly, the rate of recovery can be expressed as the reciprocal of the expected length of time between the infected state and the recovered state; let's call this expected length of time $E(T_I)$.

By this reasoning, if the average age of infection is 10 years old, then the average incidence rate of disease per person would be 1/10 $years^{-1}$. The average incidence rate per person per unit time is typically called the *force of infection*, and often abbreviated with the Greek letter lambda, λ.

Analogously, if the duration of disease is typically 6 months (0.5 years), then the rate of recovery from the disease per infected person per year is $1/(0.5) = 2$ $years^{-1}$. Let's call this rate ν.

There are two more key rates that haven't been drawn into our diagram, but that we'll need to account for in many infectious disease models. One is the birth rate, and the other is the death rate. If the population we are simulating is demographically stable—that is, if it's not rising or declining in population size—and the disease is not lethal, then the birth and death rates will be equal. If the average lifetime in the population (let's call it $E(T_L)$) is 75 years old, then if we assume an exponential distribution of life expectancy as per Equation 9.1, the rate of death in the

assume an exponential distribution of life expectancy as per Equation 9.1, the rate of death in the population per person would 1/75 years^{-1}, which in a demographically stable population will also be the rate of birth. We'll call this rate μ.

Now that we have all rates of flow that we need between states, we can write down the equations to describe our model. The equations are easier to derive if we first create a table in which we itemize every rate of flow that enters or leaves each state of our model. We can go through the model state by state to write down the people who enter the state and the people who leave the state, as shown in Table 5.1. For the purposes of the Table and equations, we can specify that the susceptible state is letter *S*, the infected state letter *I*, and the recovered state letter *R*. Furthermore, letter *N* is the total population size (equal to $S + I + R$).

[INSERT TABLE 5.1 HERE]

Translating the table into equation format, we can arrive at Equations 9.6 through 9.8, where *t* reflects time:

[Equation 9.6]

[Equation 9.7]

[Equation 9.8]

We can quickly check that we have included all of the important rates of disease by adding each of the equations together. Since we have assumed a constant population size, the sum of the right-hand sides of the equations should equal zero, meaning that there is no change in the overall population size. We can the expression $\mu*(N-(S+I+R))$, but since $N = (S+I+R)$, the overall expression does in fact add to zero.

As with our Markov models, we can first solve our infectious disease model at a steady state. Under typical steady-state conditions, meaning that the infectious disease roughly infects the same number of people each year, the prevalence of disease is equal to the incidence of disease, multiplied by the duration of disease. To understand this relationship, imagine the analogy of a restaurant: if customers in a restaurant come to eat at a rate of 10 per hour (the incidence of customers) and each stay for 2 hours (the duration of being a customer), then at any given time, we could reason that the prevalence of customers would be an old batch of 10 customers from the previous hour (who are now in their 2nd hour of eating), and a new batch of 10 from the current hour (who are still in their 1st hour of eating), for a total of 20 customers.

Analogously, we can derive an expression to estimate the steady state population of infected people in a community. First, we can estimate the prevalence of susceptible people in the stable population. This prevalence will be the "incidence" of being susceptible (i.e., births, μN) multiplied by the "duration" of being in the susceptible state (i.e., the length of time before either dying at rate μ or getting infected at rate λ, which are the only two ways to leave the susceptible state in our model). The expected duration of being in the susceptible state will be E[minimum(T_L or T_S)], since a person will either die first or get infected first, if moving out of the susceptible state. Making use of our survivor function (Equation 9.1):

$$E(\min[T_L, T_S]) = \int_0^\infty e^{-(\mu+\lambda)t}dt = \frac{1}{\mu+\lambda}$$

[Equation 9.9]

So the number of susceptible people will be the incidence of susceptible people μN times the duration of susceptibility $1/(\mu+\lambda)$, which produces Equation 9.10:

[Equation 9.10]

Similarly, we can estimate the prevalence of infected people, which will be the incidence of infection (λ times the susceptible population given by equation 9.10), and the duration of infection ($1/(\nu+\mu)$, by the same logic as above). Hence, the number of infected people will be expressed by Equation 9.11:

[Equation 9.11]

Equations 9.10 and 9.11 provide us with a useful strategy to estimate prevalence rates for stable diseases in a population that do not wildly fluctuate in incidence each year (*endemic*, rather than epidemic, diseases). For example, if we have a population of 100,000 people with an average life expectancy of 75 years, a force of infection of the disease of interest of 1/10 years^{-1}, and a rate of recovery of 2 years^{-1}, we can calculate the number of susceptible people in the context of this non-fatal, endemic disease at steady state:

[Equation 9.12]

[Equation 9.12]

Around 11,765 people would be expected to remain susceptible to the disease at steady state. Analogously, the number of infected people would be:

[Equation 9.13]

Finally, the number of people who have recovered would be the population size *N* minus the number of susceptible people, minus the number of infected people, which is 87,651.

<2>The Law of Mass Action

While it may be of interest that we can calculate these quantities just using a piece of paper and a pencil, we want to understand how to control diseases that are not at a simple endemic steady state condition. We must address infectious diseases that are fatal. We can do so by modifying Equations 9.6 through 9.8 by adding an additional rate of death from the infected state, to indicate the extra death caused by the disease. This might change the demography of the population, if the additional deaths from the disease are not exactly compensated by new births.

Furthermore, we should modify our equations to understand how disease may start out as rare but emerge into epidemics. How can we account for the fact that, by definition, infectious diseases often do not have constant rates of infection, but rather have changing rates of infection based on the number of infectious people in the population? If there are a lot of infectious people in the population, the risk of infection will be high. If a large number of infectious people are cured of their disease, then the risk of infection will go down.

To solve this problem, we can allude to the challenge we faced when we encountered a changing demand for peeps; we derived Equation 9.5 to describe how the rate of demand might change over time. Now we can use that equation to describe how the rate of infection can change over time, to deal with epidemic diseases for which the infection rate is dynamic.

The most common approach to introduce a dynamic rate of infection is called the *Law of Mass Action*. The Law of Mass Action states that in a homogeneously-mixed population (i.e., a population in which people are circulating around and of generally equivalent risk of being infected with the disease), then the incidence rate of the disease will be proportional to the population of infectious people (e.g., the infectious people will be dispersed throughout the community and collide with susceptible people):

[Equation 9.14]

For example, suppose that we have a disease for which the force of infection is 1/10 $years^{-1}$ and the number of people infected currently is 584. We would calculate =(1/10)/584 = 0.000171. It is typically the case that is a small number, because it expresses the risk of disease upon one contact between an infectious and a susceptible person per unit time, which should be small.

In equation 9.14, $\lambda(t)$ is our hazard function for the incidence rate, which changes over time depending on how many infectious $I(t)$ people there are at the time. The constant β is a transmission rate per person per unit time. It tells us how likely it is that a contact between one susceptible person and one infectious person will result in transmission of the disease in question. As a result, it doesn't change for a specific disease in a specific community, but indicates the "contagiousness" of the specific disease being studied in a particular community.

Using the Law of Mass Action, we can re-write our series of differential equations (equation 1.6) by replacing all of the βs with βI's, to keep track of an epidemic disease whose infection rate changes:

[Equation 9.15]

[Equation 9.16]

[Equation 9.17]

Of course, we can still use these equations for an endemic disease at steady-state, for which I(t) will be (by definition) constant, such that the force of infection λ will be constant at all times *t*.

It is common to make some optional modifications to Equation 9.15 through 9.17, particularly by replacing β with two variables multiplied together: the probability of contact between infectious and susceptible people, and the probability that this contact will result in disease transmission. This is useful, for example, when modeling some sexually-transmitted diseases, as we will see later in this chapter. Alternatively, it is possible to change β to equal

□I(t)/N, so that □ is multiplied by the fraction of the population that is infectious. The expression scales up the value of □/□□based on the population size.

<2>The reproductive number

The *basic reproductive number*, or R_0 (commonly pronounced "are-naught", due to its British origins) is the mean number of secondary infections a single infected person will cause in a population with no immunity to the disease, and in the absence of interventions to control the infection. If we imagine an island of people who have never been affected by the disease, then R_0 is the number of people who would be infected by one infected person who lands on the island (a fully susceptible population), before the infected person recovers or dies (is no longer infectious). Figure 5.2 illustrates a disease with an R_0 of 3.

[INSERT FIGURE 5.2 HERE]

It is important not to confuse R_0 with the *effective reproductive number* (R or R_e, depending on the author), which is the number of secondary cases generated by an infectious case once the epidemic is underway (e.g., once the population is no longer fully susceptible, but there are some immune persons or some interventions have been introduced such as medical treatment). Of course, we're interested in driving down the effective reproductive number by creating public health programs, but the basic reproductive number is a fundamental property of the disease spread in a given population.

We can reason that if the R_0 of a disease is greater than 1, then one infected person will produce more than one infected person in the next generation of infections on average, hence the disease will spread. Conversely, if R_0 is less than 1, then one infected person will produce less than one infected person in the next generation of infections on average, hence the epidemic will eventually burn out.

For epidemic diseases that initially expand, there is some point at which the number of susceptible people will be so few that infected people will not be able to transfer the disease onto as many susceptible people (most people who are in contact with the infected person are either recovered or infected). Hence, the disease transmission process will reach a point when the effective R will be 1, meaning there is some population size S such that:

[Equation 9.18]

Hence:

[Equation 9.19]

The point at which the effective R will be equal to 1 is effectively the point we solved for earlier in the endemic steady state disease situation, in which the prevalence of being in a given state is equal to the incidence of moving into the state times the duration of being in the state. The "incidence" of being in the population is the birth rate □ times the population that can give birth N, and the "duration" of being in the population is the life expectancy $E(T_L)$. Similarly, the "incidence" of being susceptible is the birth rate □ times the population that can give birth N, and the "duration" of being susceptible is E[minimum(T_L or T_S)], since a person will either die first or get infected first, if moving out of the susceptible state. Therefore, Equation 9.19 becomes Equation 9.20:

[Equation 9.20]

The approximation in the midst of this derivation is because T_S is far smaller than T_L in the vast, vast majority of cases (since life expectancy would be longer than the time to get disease).

We can simplify our expression for R_0 further by remembering that □/□□□□/□I, and using equation 9.13:

[Equation 9.21] =

After some simplification and rearrangement, we can substitute equation 9.21 into equation 9.20 and end up with this very well-known formula for R_0:

[Equation 9.22] $$R_0 = 1 + \frac{\lambda}{\mu} = \frac{\beta N}{\nu + \mu} \approx \frac{\beta N}{\nu}$$

The last approximation is because ν is much larger than □ usually.

What we've accomplished here is to use the endemic steady-state expressions to derive a

What we've accomplished here is to use the endemic steady state expressions to derive a generalizable expression for R_0. This can greatly help us to understand the dangers of a disease and the conditions under which we might be able to control the disease. For example, suppose we have a population of 100,000 people and a disease for which =0.000171 and v=2 years^{-1}, then using Equation 9.22, R_0 = (0.000171*100000)/2 = 8.6. Each infected person will on average infect 8.6 others in a fully susceptible population.

<2>Implementing the Kermack-McKendrick model in R

An example of *R* code to program the model is provided on the textbook website. To program the Kermack-McKendrick model in *R*, we specify the same key elements that we specified for the Excel model: the parameters that will go into our equations, how long we want to run the simulation, the initial conditions, and the equations themselves.

First, we input our parameters:

```
N = 100000
mu = 1/75
beta = 0.000171
v = 2
```

Second, we specify how long we want to run the model, and with what time steps.

```
time = 5
dt = 0.01
```

Third, we input our initial conditions. We'll also create some vectors named after each state, which we'll later expand for future time points, but which for now will only contain the initial conditions:

```
S = 99999
I = 1
R = 0

Svec = S
Ivec = I
Rvec = R
```

Finally, we write our equations and insert them into two "for loops" to have *R* update over each period of time (first for loop) and across all small time steps (second for loop). In each loop, we concatenate (add on) the current value of the state to the list of previous values, expanding the vector with each time step across all simulated time points. Note that there are more efficient ways to code this using the library 'deSolve', but this way is more intuitive for learning purposes:

```
for (i in 1:time){
 for (i in 1:(1/dt)){
   S = S + mu*N*dt - beta*S*I*dt - mu*S*dt
   I = I + beta*S*I*dt - v*I*dt - mu*I*dt
   R = R + v*I*dt - mu*R*dt
   Svec = c(Svec, S)
   Ivec = c(Ivec, I)
   Rvec = c(Rvec, R)
 }
}
```

Now we can plot the result; here, we use the Quick-R website (http://www.statmethods.net/graphs/line.html) to remind ourselves of how to color the lines blue `(col="blue")`, choose a smooth line `(type="l")`, and label our axes `(xlab="time steps", ylab="Pop")`. We choose one vector to start the plot with the plot command, then add the other two lines using the lines command:

```
plot(Svec,col="blue",type="l",xlab="time steps",ylab="Pop")
lines(Ivec,col="red")
lines(Rvec,col="green")
```

The Quick-R site also reminds us how to add a legend by specifying the x-axis location for the left corner of the legend box, the y-axis location for the top of the legend box, the labels for the legend, the type of lines we want in the legend, and the colors.:

```
legend(400,80000,c("S","I","R"),lty=c(1,1,1),col=c("blue","red","green"))
```

The plot commands produce Figure 5.6:

[INSERT FIGURE 5.6 HERE]

We can use the power of *R* to extend the model to do something we simply don't have the ability to do in Excel: a large-scale uncertainty analysis. Suppose that we don't know exactly what the value of the input parameters are. We might know, for example, that β = 0.000171 on average, but has a standard deviation of 0.00001, and v = 2 on average, but has a standard deviation of 0.1. How much would this affect the number of people infected in the first year of the epidemic?

We can first create a vector to keep track of how many total people get infected over the one year simulation, by adding a vector Totinf = 0 before the "for loop" and adding within the for loop the code:

```
Totvec = c(Totvec, beta*S*I*dt)
```

This code keeps track of how many people are newly infected in each time step. Hence, the sum from time 0 to time 1/dt will be the total number of people infected over the course of the first year of the epidemic:

```
> sum(Totvec)
[1] 101842.4
```

Next, we can make our model *stochastic* instead of *deterministic*. That is, rather that having the value of the parameters pre-determined at set values, we can have the model incorporate parameters that are stochastic, or chosen from random distributions we specify. In this case, our initial parameters will now be:

```
N = 100000
mu = 1/75
beta = rnorm(1,mean=0.000171,sd=0.00003)
v = rnorm(1,mean=2,sd=0.2)
```

The rnorm commands indicate that *R* should sample once from a normal distribution with means and standard deviations as specified. Now, we want to know how much the final size of the epidemic (the total number of people infected over the course of the epidemic) might vary because of the uncertainty in the parameter values. We can wrap the entire code in a larger "for loop" and run it 100 times, storing the value of sum(Totvec) in a new vector called Totveciters, which will include the value of Totvec for each of the 100 iterations:

```
Totveciters=c()
for (iters in 1:100){

N = 100000
mu = 1/75
beta = rnorm(1,mean=0.000171,sd=0.00001)
v = rnorm(1,mean=2,sd=0.1)

time = 1
dt = 0.01

S = 99999
I = 1
R = 0

Svec = S
Ivec = I
Rvec = R
Totvec = 0

for (i in 1:time){
 for (i in 1:(1/dt)){
   S = S + mu*N*dt - beta*S*I*dt - mu*S*dt
   I = I + beta*S*I*dt - v*I*dt - mu*I*dt
   R = R + v*I*dt - mu*R*dt
   Svec = c(Svec, S)
   Ivec = c(Ivec, I)
   Rvec = c(Rvec, R)
   Totvec = c(Totvec, beta*S*I*dt)
```

```
   Totvec = c(Totvec, beta^S^I^dt)
 }
}

plot(Svec,col="blue",type="l",xlab="time steps",ylab="Pop")
lines(Ivec,col="red")
lines(Rvec,col="green")
legend(400,80000,c("S","I","R"),lty=c(1,1,1),col=c("blue","red","green"))

Totveciters=c(Totveciters,sum(Totvec))
}
```

We can finally plot the histogram that tells us how much the size of our epidemic in the first year varied due to the uncertainty in our input parameters, which is shown in Figure 5.7:

[INSERT FIGURE 5.7 HERE]