Modeling an Outbreak of Influenza:

Applying a Stochastic Markov Chain Model to Predict Infections

Carly Lundgreen

December 10, 2020

Introduction

The mathematical modeling of epidemics or outbreaks of infectious disease can be difficult, and often requires many simplfying assumptions in order to apply an epidemiological model to a set of data. Predicting how a disease such as Influenza moves through a population, however, can be extremely useful in planning for needed resources and preparing properly for any future outbreaks of the disease. One such epidemiological model is the S-I-R (Susceptible, Infected, Recovered) model. This model divides the population of interest into S-I-R compartments, and individuals move between compartments/groups based on transition probabilities. We can display the model as follows:

$$\begin{pmatrix} S_t \\ I_t \\ R_t \end{pmatrix} = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ 0 & a_{32} & 1 \end{pmatrix} \begin{pmatrix} S_{t-1} \\ I_{t-1} \\ R_{t-1} \end{pmatrix}$$

Where S_t, I_t, and R_t are the probabilities of being in the respective groups at time t, and S_{t-1}, I_{t-1}, and R_{t-1} are the probabilities at time t-1. This is the Markov Chain aspect of this model–where probabilities at time t are affected only by the probabilities at time t-1. Note that we can write this model as $P_t = \mathbf{A}P_{t-1}$, where the columns of A must sum to 1 so as to keep the condition that $S_t + I_t + R_t = 1$ (the analog of this is that the sum of the numbers of individuals in each group at time t must sum to the size of the population).

We have three main goals that we would like to accomplish through this analysis. The first is to utilize the SIR model to estimate the expected number of infected individuals on a given day during an outbreak of an infectious disease, given only data about daily infection numbers. This requires estimation of the elements of $\mathbf{A}$, which are transition probabilities of staying or moving to the next group. Secondly, we will attempt to estimate how long it will likely take (in days) to reach the maximum number of infected individuals during the outbreak. Finally, we will provide a reasonable range of values for the expected number of infected individuals at time t. These results can help us understand how an infectious disease may behave if introduced to another similar population of individuals.

In order to estimate the parameters of interest, which are the elements of $\mathbf{A}$, we will use the Expectation-Maximization algorithm. This algorithm is often used as a method of imputing missing data in addition to estimating 'latent variables' as a path to parameter estimation. In our case, we must first estimate the probabiltieis of being in the susceptible and recovered groups at time t before we can estimate the parameters of interest using an optimization procedure. Once we have our S and R probabilties, we can estimate the transition probabilities of the SIR model, which will be discussed further in the next section. After we estimate these probabilities, we will utilize draws from Multinomial distribution to estimate the expected number of susceptible, infected, and recovered individuals on a given day and calculate 95% confidence intervals for the expected values of infected individuals at time t.

Methodology

Solving the right-hand side of the model outlined in the previous section gives the following system of equations:

$$\begin{cases} S_t = a_{11} S_{t-1} \\ I_t = a_{21} S_{t-1} + a_{22} I_{t-1} \\ R_t = a_{32} I_t + R_{t-1} \end{cases}$$

We want to estimate a_{21} and $a22$. We can calculate the other transition probabilities according o $a_{11} = 1 - a_{21}$ and $a_{32} = 1 - a_{22}$. Note that we only know I_t, and I_{t-1}, so we will need to consider S_t, S_{t-1}, R_t, and R_{t-1} as latent variables to proceed with estimating the transition probabilities. To do this, we apply the Expectation-Maximization algorithm. First, we select initial values of a_{21} and a_{22}. Then, we can utilize the above system of equations to estimate the probabilities associated with the S and R groups. This requires initialization of the S and R vectors. The proportion of susceptible individuals at time 0 is assumed to be 1, and the proportion of recovered individuals at time 0 is assumed to be 0.

After estimating the probabilities for the S and R groups, we will use optimization to estimate the values of a_{21} and a_{22} that minimize the following target function:

$$\sum_{i=2}^{n} (I_t - (a_{21} S_{t-1} + a_{22} I_{t-1}))^2$$

The algorithm assumes the following steps:

- Random initialization of parameters a_{21} and a_{22}
- Set $R_1 = 0$ (we assume that the outbreak begins with a proportion of 0 recovered individuals)
- Set $R_t = a_{32} I_{t-1} + R_{t-1}$ for t = 2,. . . ,n
- Set $S_t = 1 - I_t - R_t$ for all n
- Estimate a_{21} and a_{22} through optimization of the target function outlined above. Repeat steps (except for random initialization of parameters) until convergence

After convergence of the algorithm, we will re-run the Markov Process outlined above using the estimated a_{21} and a_{22} to get a new set of values for the proportion of S, I, and R individuals at time t, which results in a 14x3 matrix of probabilities. We check that each row sums to 1 to keep the condition that $S_t + I_t + R_t = 1$. Each element of the matrix is the probability of an individual being in that particular group at time t, for t=1,. . . ,n.

We are assuming a fixed and finite population size, N, so we can imagine each row of this matrix as probabilities of events within a Multinomial distribution, where n = 743 and the probabilities in each row are the Multinomial parameters $p_1, \ldots, p_i$, for i = 1,. . . ,3, where $\sum p_i = 1$. We will perform 10,000 random Multinomial draws where n=743 and $p_1, \ldots, p_i$ is a single row from probability matrix of S_t, I_t, and R_t. We can then take an average of each column of the resulting 10,000 x 3 matrix to return the expected values of S_t, I_t, and R_t and utilize the `quantile` function in R to calculate 95% confidence intervals for the expected values of the probability of infection at time t. We will repeat this for each row of the 14x3 probability matrix, resulting in a table of expected values of S_t, I_t, and R_t for all t = 1,. . . ,n.

Simulation

We conducted a simulation study to make sure that the Expectation-Maximization method will converge to two values of a_{21} and a_{22} and allow us to then utilize the Multinomial distribution to answer the research questions. The simulation first involved the creation of a function, `test`, that takes a_{21} and a_{22} as arguments and is composed of the following steps:

- Initialize values of $a_{21} = 0.5$ and $a_{22} = 0.4$. In order to generate probabilities between 0 and 1, a bit of fiddling with the initial values was necessary. We found that values close to 0.5 worked best.

- Within a function `test` that takes a_{21} and a_{22} as arguments:
- Generate n=14 Unif(0,1) variables to represent I_t. The uniform distribution on the [0,1] domain was chosen simply for the support that matches the assumption that these values are probabilities of infection at time t+1.
- Initialize $I_1 = 0.01$, a value close to 0 (we areassuming that the outbreak begins with a small fraction of the population) and $R_1 = 0$.
- Calculate R_t and S_t according to the linear equations we set up in the previous section.
- Calculate $S_t = 1 - I_t - R_t$
- Within the same function, write a function, `target`, that `optim` will minimize. This is the same target function outlined in the previous section. The function `target` will take one argument, *pars*, a vector of length 2 with respective elements a_{21} and a_{22}. These are initial guesses for parameters, and if the optimization procedure is set up correctly, the initial values shouldn't affect the estimates to which the algorithm converges.
- Set *pars* = c(0.5,0.6), arbitrary initial guesses, and run `optim(pars,target,method="BFGS")$par` to return the converged estimates.

After building the function `test` described on the previous above, repeat `test` 10,000 times using initial values $a_{21} = 0.5$ and $a_{22} = 0.53$. This is most easily done with the R function `replicate`. After completing this process, we are left with 10,000 estimates of a_{21} and a_{22}. We have plotted histograms below, with vertical lines denoting the expected value of each parameter estimate.

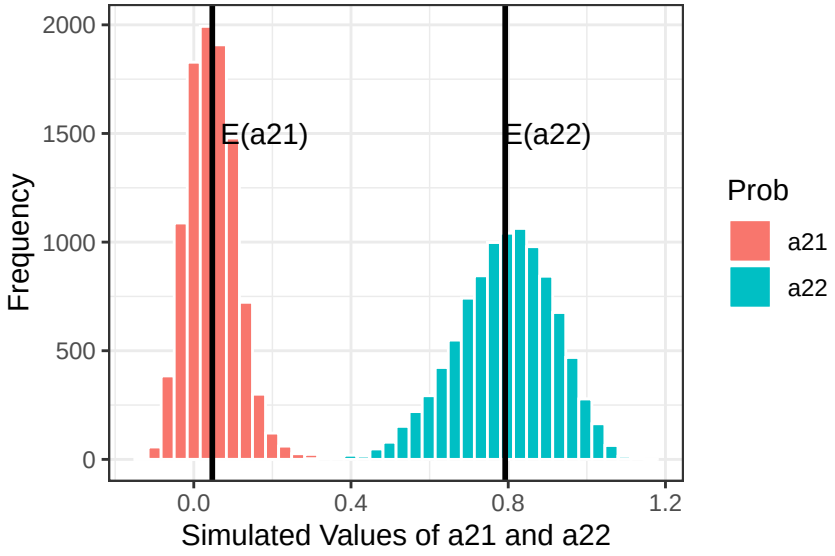

Figure 1: Histograms of 10,000 Simulated Estimates of a21 and a22

Figure 1 demonstrates that applying this methodology to a dataset will result in the algorithm's convergence to two parameter estimates, which we can then plug back into the system of equations/Markov Chain to generate probabilities for the susceptible and recovered groups.

Results

This analysis and application of the methodology will focus on a dataset that comes from a boarding school in England and reports the number of occupied sick beds on a given day during an Influenza outbreak. There are 14 observations in the dataset and the total population at the school is 743 individuals. Note that we are assuming that the number of individuals occupying sick beds equates to the number of infected individuals

on a given day, which is likely not entirely true, but will still allow us to estimate the effects of the disease outbreak in terms of the number that become infected.

A table of summary statistics for the data is displayed below in Table 1. Table 2, also below, illustrates summary statistics for the proportions of individuals infected at time t (t= 1,. . . ,14), which were calculated by dividing the infected numbers (given data) by the total number of individuals in the population (743).

Table 1: Summary Statistics for The Number of Infections Per Day

	Mean	SD	Range	Median	IQR	Min	Max
Value	110.429	106.377	3	72.5	197.25	3	291

Table 2: Summary Statistics for the Proportion of Infections Per Day

	Mean	SD	Range	Median	IQR	Min	Max
Value	0.149	0.143	0.004	0.098	0.265	0.004	0.392

Based on Table 1, an average of about 110 people became infected over the course of the outbreak, beginning with 3 individuals at time 0 (day 1). At the peak of the outbreak, 291 individuals were sick. We can also see that at the peak, the proportion of infected individuals was 0.392, according to Table 2.

To apply our methodology, we re-run simulation process a single time using the same initial values, but instead of generating n=14 Unif(0,1) variables to represent I_t, we will use the given data (probabilities of infection at time t, not the counts). We want to return both the parameter estimates for a_{21} and a_{22} as well as the estimated probabilities of S_t and R_t, so a slight modification of the `test` function may be needed to return the estimated probabilities in addition to the estimates. After applying this methodology to the data, the `optim` function converged to $a_{21} = 0.087$ and $a_{22} = 0.826$. Solving for the other transition probabilities gives the following Markov Chain SIR model:

$$\begin{pmatrix} S_t \\ I_t \\ R_t \end{pmatrix} = \begin{pmatrix} 0.913 & 0 & 0 \\ 0.087 & 0.826 & 0 \\ 0 & 0.174 & 1 \end{pmatrix} \begin{pmatrix} S_{t-1} \\ I_{t-1} \\ R_{t-1} \end{pmatrix}$$

$$\begin{cases} S_t = 0.913 S_{t-1} \\ I_t = 0.087 S_{t-1} + 0.826 I_{t-1} \\ R_t = 0.174 I_t + R_{t-1} \end{cases}$$

The following plot, Figure 2, illustrates the model by assigning weights to lines that move between the three stages–representing the progress of the outbreak in terms of time t and time t+1.

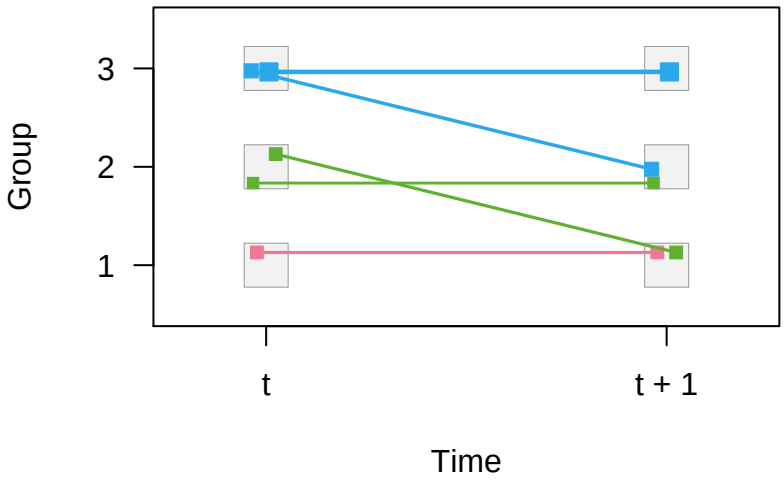

Figure 2: Illustrative Plot of Transition Matrix Probabilities and Movement Between Compartments

Figure 2 refers to the S, I, and R groups as groups 1, 2, and 3, respectively. Each of the lines are color coded to represent a transition to the corresponding group. The lines are also weighted by the five transition probabilities from matrix **A**. This plot allows us to infer that an individual that is in the susceptible group at time t is very likely to still be susceptible at time t+1. Similarly, if an individual is infected at time t, it is very likely that at time t+1 they will still be infected. The probability of moving between the susceptible and infected group is relatively small, and combined with the probability of staying in the infected group, we can calculate I_t, or the probability of infection at time t. The blue line at the top of this plot is weighted with a 1, indicating the assumption that once an individual has recovered from infection, they cannot re-enter the susceptible group.

In addition to understanding the transition probabilities, the results of our algorithm allow us to create the following plot, which illustrates the the progress of the Influenza outbreak over time with respect to the S, I, and R groups.

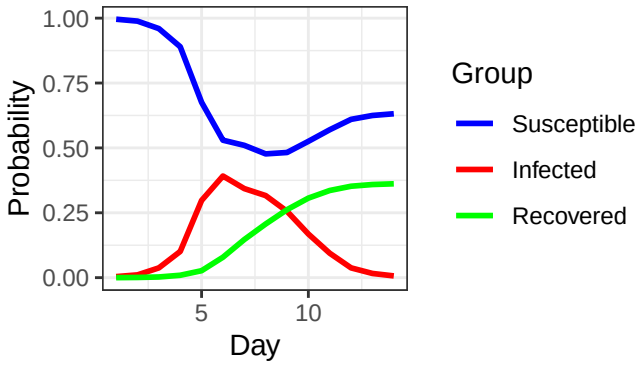

Figure 3: Line Plot Illustrating the Progress of Outbreak Over Time

We can see from Figure 3 that the peak of the outbreak's infections appears just after day 5. Also, it appears

that at the end of outbreak, individuals in the population were more likely to still be in the susceptible group than infected or recovered. This indicates that the illness did not move through a majority of the population. Therefore, even if infection leads to immunity from this particular Influenza strain, a re-introduction of the disease into the same population would still lead to a relatively large number of infections.

Now that we have estimated the elements of the **A** matrix in the SIR model using only the number of infected individuals at time t, we can now calculate expected values of infection (as well as susceptibility and recovery) at time t using a Multinomial distributional application, the process of which was outlined in the Methodology section. Using the S, I, and R probabilities (a 14x3 matrix), we used each row of the matrix as probabilities of events within a Multinomial distribution with n=743 "trials". This method would not apply if the population was not fixed at a single value. A constraint of the Multinomial distribution parameters $p_1, \ldots, pi$ is that they must sum to 1, which is inherently a condition of the SIR model where $S_t + I_t + R_t = 1$. After performing 10,000 draws from a multinomial distribution, applied to each row of the 14x3 matrix of probabilities, we generate the following table. Note that we have converted back from probabilities to counts by multiplying by 743.

Table 3: Expected Values at Time t with 95% CI for Infected Only

Day	S	I	R	2.5%	97.5%
1	739.99	3.01	0.00	0	7
2	734.49	7.99	0.53	3	14
3	713.06	28.03	1.91	18	39
4	661.28	74.88	6.84	59	91
5	502.16	221.00	19.84	197	245
6	393.65	291.08	58.27	265	317
7	378.49	255.20	109.31	229	281
8	354.43	234.75	153.82	210	260
9	358.30	190.14	194.56	167	214
10	390.36	125.03	227.62	105	145
11	423.36	70.06	249.58	55	86
12	453.21	27.98	261.81	18	39
13	464.52	11.99	266.49	6	19
14	469.15	5.03	268.82	1	10

We can see that, for 10,000 random draws, the expected number of individuals that are infected is the highest on day 6 of the outbreak, with a 95% confidence interval of (265,317). We could use this information to suggest a preparation of approximately 300 beds if the disease is re-introduced to the same population. We can see that on day 14 of the outbreak, about 470 individuals were still susceptible. This indicates that the disease moved through approximately 743-470=273 individuals over the course of two weeks.

Conclusion

The SIR model is a very useful method of modeling the spread of infectious disease. Estimating the transition probabilities through an application of the Expectation-Maximization algorithm allowed us to then employ the Multinomial distribution to estimate the expected number of individuals infected at time t, along with 95% confidence intervals for that expected value. We also note that there are a few limitations of the SIR model in the form of multiple simplifying assumptions. For example, the model assumes that there is homogeneous mixing of the infected and susceptible populations, which implies that individuals associate/make contact with each other at random. This assumption does not feel very reasonable at a school where children have close associations or friend groups.

The model also assumes that that rates at which individuals enter and leave groups do not change. We also assumed that the total population stays constant. In the case of a boarding school, it is reasonable that

no one is enrolling in or dropping out of school over the course of the outbreak. It is also assumed that recovered individuals cannot re-enter the susceptible group, which is always dependent upon the particular illness and immunities within the population. There are also other more complicated models that we did not consider and perhaps would fit these data better. Overall, however, the SIR model can be a very effective way to predict the movement of an infectious disease through a population, especially in the case of a closed population with relatively controlled conditions such as we saw in the English boarding school.