Modeling an Influenza Epidemic

Applying a Stochastic Markov Chain Model to Predict Infections

Carly Lundgreen

Background and Research Questions

Background:

- This dataset 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
- We are assuming that the number of individuals occupying sick beds equates to the number of infected individuals on a given day

Goals of the Analysis:

- Use a stochastic S-I-R model to estimate the expected number of infected individuals on a given day during the epidemic
- Infer how long it will likely take to reach the maximum number of infected individuals
- Provide a reasonable range of values for that maximum

Data

Day	# Sick	Probability
1	3	0.004
2	8	0.011
3	28	0.038
4	75	0.101
5	221	0.297
6	291	0.392
7	255	0.343
8	235	0.316
9	190	0.256
10	125	0.168
11	70	0.094
12	28	0.038
13	12	0.016
14	5	0.007

SIR Model

The SIR (Susceptible, Infected, Recovered) model is a method to measure the spread of infectious disease over time. The model can be written as follows:

$$\begin{pmatrix} P_{S\,t} \\ P_{I\,t} \\ P_{R\,t} \end{pmatrix} = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ 0 & a_{32} & 1 \end{pmatrix} \begin{pmatrix} P_{S\,t-1} \\ P_{I\,t-1} \\ P_{R\,t-1} \end{pmatrix}$$

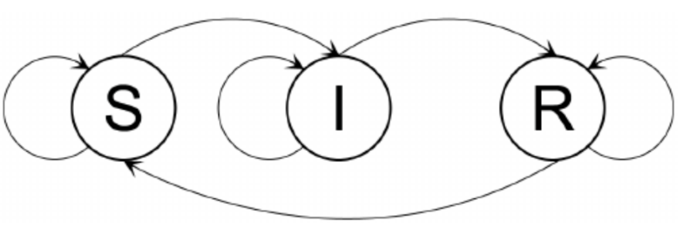
which can be simplified to:

$$P_t = \mathbf{A} P_{t-1}$$

and has the property that $P_{R\,t} + P_{S\,t} + P_{I\,t} = 1$ (i.e. columns of $\mathbf{A}$ sum to 1)

SIR Model, Ctd.

The elements of **A** are transition probabilities of moving to the next group or staying in one's current group. The Markov chain of these transition probabilities in the SIR model can be displayed as follows:



What are we trying to estimate?

Solving the right-hand side of the model gives the following system of equations:

$$\begin{cases} P_{S\,t} = a_{11} P_{S\,t-1} \\ P_{I\,t} = a_{21} P_{S\,t-1} + a_{22} P_{I\,t-1} \\ P_{R\,t} = a_{32} P_{I\,t-1} + P_{R\,t-1} \end{cases}$$

Want to estimate the elements of **A**, but we only know $P_{I\,t}$ and $P_{I\,t-1}$. We need to consider $P_{S\,t}$, $P_{S\,t-1}$, $P_{R\,t}$, and $P_{R\,t-1}$ as latent variables to proceed.

Strategy

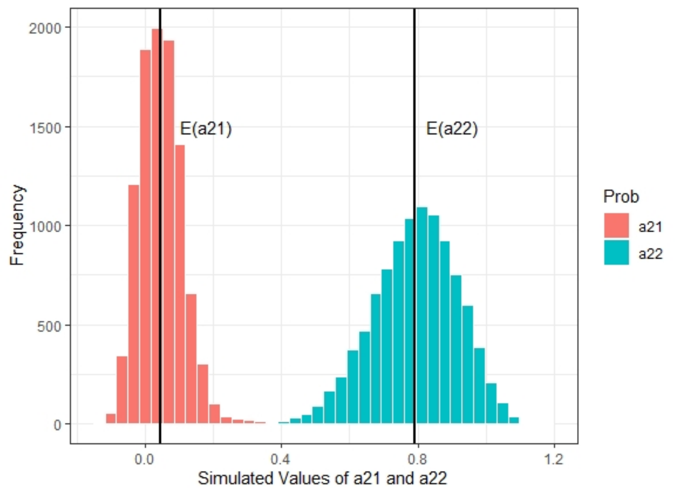
1. Use system of linear equations to find expected values of $P_{S\,t}$ and $P_{R\,t}$ after choosing initial values of a_{21} and a_{22}
2. Minimize $\sum_{t=2}^{n}(P_{I\,t} - (a_{21}P_{S\,t-1} + a_{22}P_{I\,t-1}))^2$ using *optim* to get estimates of a_{21} and a_{22}
   - Steps 1 and 2 are an application of the Expectation-Maximization algorithm
3. Re-run Markov Chain process using the estimated a_{21} and a_{22} to get a new set of infection probabilities
4. Use the Multinomial distribution to generate 10,000 vectors of length 3 representing expected value of S, I, and R individuals at time t. Calculate 95% intervals for $E(I_t)$ using *quantile*

Simulation

- Initialize values of $a_{21} = 0.5$ and $a_{22} = 0.4$
- Within a function (*test*) that takes a_{21} and a_{22} as arguments:
    - Generate n=14 Unif(0,1) variables to represent $P_{I\,t}$
    - Initialize $P_{I\,1} = 0.01$, $P_{R\,1} = 0$
    - Calculate $P_{R\,t}$ and $P_{S\,t}$ according to the linear equations we set up
    - Calculate $P_{S\,t} = 1 - P_{I\,t} - P_{R\,t}$
    - Within the same function, write a function, *target*, that *optim* will minimize: $\sum_{t=2}^{n}(P_{I\,t} - (a_{21}P_{S\,t-1} + a_{22}P_{I\,t-1}))^2$
        - *target* has one argument *pars*, a vector of length 2 with elements a_{21} and a_{22} (initial guesses for parameters)
    - Set $pars = c(0.5, 0.6)$, run optim(pars,target,method="BFGS"), return estimates

Simulation Results

After building the function *test* described on the previous slide, repeat *test* Nsim=10,000 times using initial values $a_{21} = 0.5$ and $a_{22} = 0.53$ (but any initial values will lead to convergence)



Application to Data

Re-run simulation process a single time using the same initial values, but instead of generating n=14 Unif(0,1) variables for P_{It} use the given data (probabilities, not counts)

We want to return both the parameter estimates for a_{21} and a_{22} as well as the estimated probabilities of P_{St} and P_{Rt}, so a slight modification of the *test* function may be needed.

After applying the method to the data, the *optim* function converged to:

a_{21}	a_{22}
0.087	0.826

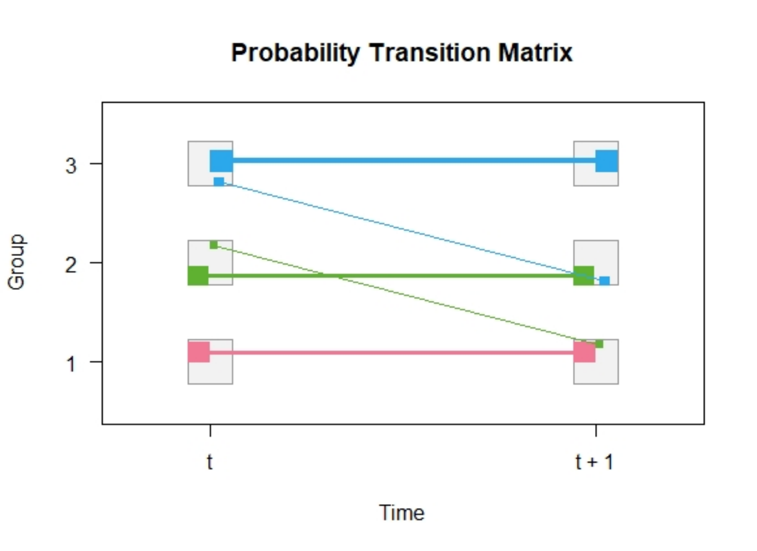
Transition/Stochastic Matrix

We can rebuild the **A** matrix after estimating a_{21} and a_{22}:

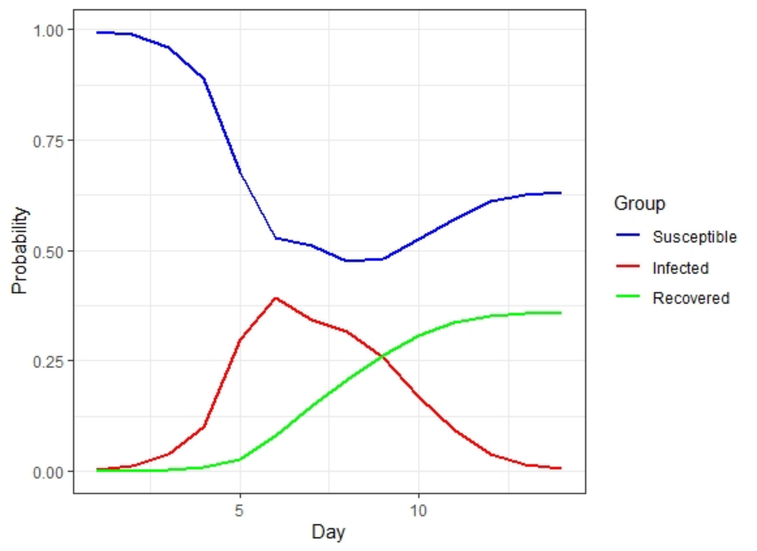
$$\begin{pmatrix} P_{S\,t} \\ P_{I\,t} \\ P_{R\,t} \end{pmatrix} = \begin{pmatrix} 0.913 & 0 & 0 \\ 0.087 & 0.826 & 0 \\ 0 & 0.174 & 1 \end{pmatrix} \begin{pmatrix} P_{S\,t-1} \\ P_{I\,t-1} \\ P_{R\,t-1} \end{pmatrix}$$

$$\begin{cases} P_{S\,t} = 0.913 P_{S\,t-1} \\ P_{I\,t} = 0.087 P_{S\,t-1} + 0.826 P_{I\,t-1} \\ P_{R\,t} = 0.174 P_{I\,t-1} + P_{R\,t-1} \end{cases}$$

Transition Matrix Diagram



Epidemic Progress Over Time



Apply Multinomial Distribution

Re-run Markov Process using estimated a_{21} and a_{22} to get a new set of values for the number of S, I, and R individuals at time t.

We now have a 14x3 matrix of probabilities, checking that each row sums to 1 to keep the condition that $P_{St} + P_{It} + P_{Rt} = 1$. Each element of the matrix is the probability of an individual being in that particular group at time t, t = 1, . . . , 14.

Because we know N = 743 (the size of the population), 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 parameters $p_1, \ldots, p_i$, for i = 1,. . . ,3, where $\Sigma p_i = 1$.

Apply Multinomial Distribution, ctd.

Draw from Multinomial Distribution:

- Perform 10,000 random Multinomial draws where n=743 and $p_1, \ldots, p_i$ is a single row from probability matrix of P_{St}, P_{It}, and P_{Rt}
- Return the mean of each column and 2.5% and 97.5% quantiles on the second column, P_{It}, to calculate a 95% interval for the expected number of infected individuals at time t
- Repeat for each row of probability matrix

Results

The procedure results in the following table. 95% confidence interval bounds refer to $P_{I\,t}$ only.

Day	S	I	R	2.5%	97.5%
1	740.00	3.00	0.000	0	7
2	734.46	8.02	0.53	13	14
3	713.02	28.05	1.93	18	39
4	661.06	75.13	6.81	59	92
5	502.21	220.87	19.93	196	246
6	383.31	291.34	58.34	266	317
7	378.79	255.05	109.16	230	281
8	354.44	234.99	153.57	210	260
9	358.33	190.05	194.61	167	214
10	390.34	124.86	227.80	105	145
11	423.25	70.12	249.63	55	86
12	453.09	28.07	261.84	18	39
13	464.22	11.96	266.82	6	19
14	469.30	4.97	268.73	1	10

Limitations

The SIR Model makes quite a few simplifying assumptions in order to create such a simple deterministic model:

- Homogeneous mixing of the infected and susceptible populations
    - (i.e. individuals associate/make contact with each other at random. This doesn't feel very reasonable at a school where kids have close associations/friend groups)
- The rates at which individuals enter and leave groups do not change
- Total population is constant
    - No one is enrolling in/dropping out of the school during the epidemic, which feels reasonable in a closed population like a boarding school
- Recovered individuals cannot re-enter susceptible group
- There are other more complicated compartmental models that we did not consider, and maybe a compartmental model isn't even the correct choice