SIR Model

Jackson Curtis

Brigham Young University

Abstract. This project explores parameter estimation of the SIR (Susceptible-Infected-Recovered) Markov chain epidemiology model. I derive the likelihood estimator for the special case where only one of the three states is fully observed. We apply the model to a data set tracking the spread of a disease through an English boarding school.

1 Introduction

The Susceptible-Infected-Recovered model (SIR) is used in epidemiology to describe the spread of a disease throughout the population. Each person in the population is expected to transition from susceptible to ill to recovered with equal probabilities. The following represents our model of interest

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

The first vector represents the expected proportion of the population in each of the categories and is a function of the transition matrix and the probabilities from the time period immediately preceding. The matrix with $a_{ij}s$ is the transition matrix for the Markov chain. According to Markov chain assumptions, a person has the probability of transitioning from state i to state j in one time period (regardless of how they got to state i) with the probability given in column i, row j of the matrix. For our case we assign 0 probability of jumping from susceptible to recovered, and a probability of 1 for staying recovered once reaching the recovered state. An assumption we will make throughout the paper is that our initial state, P_0 is that one person was infected and the rest were susceptible ($P_{I,0} = 1/n$ and $P_{S,0} = (n-1)/n$).

Our data set of interest is count data of infected students in an English boarding school over fourteen consecutive time periods. An interesting feature that will complicate our analysis is that only the population size (743) and the number sick each day are given. We do not know how many recovered students there were or how many never got sick. We will derive our likelihood to account for this unobserved information.

Because the columns of the transition matrix must sum to one, only two parameters vary without restriction while the other two are fixed. For ease of interpretation, we will model a_{12} as the probability of catching the disease, and a_{32} as the probability of recovering. These two will be the parameters we obtain through maximum likelihood estimation.

2 Deriving the Likelihood

Because our data is a sequence of observations in time, our data points are not independent or identically distributed. Instead we can write the likelihood as:

$$f(x_1, x_2, ..., x_n | a_{12}, a_{32}) = f(x_1|a_{12}, a_{32}) f(x_2|x_1, a_{12}, a_{32}) ... f(x_n|x_{n-1}, a_{12}, a_{32}) \tag{2}$$

where x_i is the number observed infected at time period i. Each observation is dependent only on the observations that proceeded it.

To deal with the fact that only those in the second state are observed and reported, we can calculate the probabilities of the number of individuals in each of the either two states. By doing this we can model each part of the likelihood as:

$$f(x_i|x_{i-1}, a_{12}, a_{32}) = \sum_{j=0}^{n} f(x_i|x_{i-1}, a_{12}, a_{32}, R_i = j) \cdot Pr(R_i = j) \tag{3}$$

Where R_i is a random variable representing the number of people recovered at the start of time period i.

To demonstrate this, consider the first case, R_1. In R_1, $Pr(R_1 = 0) = 1$ since no one has had any time to recover. In day two, the original infected person could have recovered (with probability a_{32}) or not (with probability a_{22}), therefore $Pr(R_2 = 0) = a_{22}$ and $Pr(R_2 = 1) = a_{32}$. By summing over the non-zero probabilities in R_i, we can calculate the full probability of x_i without knowing how many are recovered or infected. While future states get much more complicated, we can continue to calculate R_i as:

$$Pr(R_i = r + h) = \sum_{j=0}^{x_{i-1}} Pr(H_i = j|R_{i-1} = r - j) \cdot Pr(R_{i-1} = r - j) \tag{4}$$

Where H_i represents the number of people who recovered on day i, a number that ranges between 0 and the number of people who were sick the observation before x_{i-1}.

Renormalizing R_i

In certain scenarios an observation, x_i, may invalidate what we knew about the probability distribution of R_i. Consider the simple case of a population of ten and the observed data of (5, 6, 4, 0, 1, 0). Calculating R_5 as derived above, we would conclude that $Pr(R_5 = 10) \neq 0$. In other words, from the first four observations we would conclude that it is possible that all ten individuals have already recovered. However, x_5 invalidates this. To compensate when new information invalidates old probabilities, we renormalize the densities so that they still sum to one. This is akin to rolling a dice without observing the outcome

and assigning 1/6 to each probability. If you obtain information that the roll was not a 1, the proper way to update the probabilities on the other five possibilities is to divide by the sum of the remaining probabilities (5/6). I employ a similar strategy here.

Likelihood Given R

Now that we have a way to track R_i, we can focus on the first piece of (3). If we know R, we can write a closed form of the equation as follows:

$$f(x_i|x_{i-1}, a_{12}, a_{32}, r) = \sum_{h=0}^{x_{i-1}} \binom{x_{i-1}}{h} a_{32}^{h}(1-a_{32})^{x_{i-1}-h} \binom{n - x_{i-1} - r}{x_i - (x_{i-1} - h)} a_{12}^{x_i-(x_{i-1}-h)}(1-a_{12})^{n-r-x_i} I(x) \tag{5}$$

$$\text{where } I(x) = \begin{cases} 1 & \text{for } x_{i-1} - h \leq x_i \leq n - r \\ 0 & \text{otherwise} \end{cases}$$

This equation breaks down into sums of two binomial densities, one for how many people recover on time period i and one for how many people get ill on time period i. If the number of people who would need to get sick in order to observe x_i exceeds the number of people who have yet to get sick, the contribution to the probability is 0, which is expressed in the indicator function.

Computational Details

The resulting likelihood is complex. A triple summation is required to evaluate the likelihood–one for the time periods, one for the number of individuals recovered before day i, and one for the number recovered on day i. The partial derivatives required to obtain the maximum likelihood estimators analytically are intractable, so we will use the Nelder-Mead optimization algorithm to maximize the log-likelihood. This requires many evaluations of the log-likelihood. To facilitate faster evaluation, the log-likelihood function was coded in C++.

3 Evaluating Estimator Performance

After obtaining a method of calculating maximum likelihood estimates, I tested how they perform at estimating data where the true parameters were known. I was particularly interested in the bias and the variance of the estimates produced from the MLE for different sample sizes and observation times. In order to test this I ran a simulation study where I generated the data using a known Markov chain transition matrix and reported the observed infected at each time period. I then used my likelihood function to estimate the parameters. The following

Pop.	No obs.	a_{12}	a_{32}	95% CI on $\hat{a}_{12}$	95% CI on $\hat{a}_{32}$	$\hat{a}_{12}$ sd	$\hat{a}_{32}$ sd
743	14	0.0848	0.3420	(0.0845, 0.0857)	(0.341, 0.345)	0.0070	0.0212
500	14	0.02	0.3	(0.020, 0.021)	(0.305, 0.319)	0.0046	0.0766
473	14	0.0848	0.3420	(0.084, 0.086)	(0.340, 0.345)	0.0087	0.0271
200	6	0.1	.3	(0.101, 0.104)	(0.301, 0.314)	0.0174	0.0760
100	14	0.2	0.7	(0.202, 0.206)	(0.700, 0.708)	0.0269	0.0482
100	14	0.7	0.2	(0.700, 0.707)	(0.222, 0.225)	0.0435	0.0214
100	14	0.5	0.5	(0.505, 0.513)	(0.534, 0.537)	0.0450	0.0405
50	20	0.1	0.3	(0.099, 0.104)	(0.282, 0.291)	0.0270	0.0500
20	8	0.4	0.3	(0.405, 0.422)	(0.304, 0.315)	0.0966	0.0668

Table 1: Simulation results for 500 simulations of each scenario. Green means parameter is within interval, red if below interval, blue if above.

tables summarizes my results for different combinations of the two parameters, size of population, and days observed.

The results of the simulation show that our method of estimation approximates the true values quite well. While the truth was not contained in all of the confidence intervals on the predictors, many scenarios produced unbiased estimates, and the bias appears to be small in all cases. It is unclear whether bias comes from the likelihood itself or the Nelder-Mead approximation of the maximum likelihood. In addition, Table 1 shows that the standard deviations of the estimate were small and decrease inversely with the size of our population.

4 Analysis

Boarding School Data

Having shown a reliable way to estimate the parameters, parameters for the boarding school data can be estimated. These estimates will be useful for predicting what would happen if another group was exposed to the same disease. Table 2 gives the parameter estimates. According to the estimates, a susceptible individual has about a 1/12 chance of getting sick any given day, and a sick person has a 1/3 chance of recovering any given day.

Boarding School Parameter Estimates	
$\hat{a}_{12}$	$\hat{a}_{32}$
0.0848	0.3420

Table 2: MLE estimates for boarding school data

A concern to draw attention to however is model fit. The data does not do a particularly good job of meeting model assumptions. One assumption of the model is that once a disease is present, the entire population is completely

exposed to it. However, the boarding school data seems to suggest a ramp up time where as more people get infected, the probability of getting infected goes up. Figure ??? demonstrates this graphically by simulation 200 situations where the true parameters are our MLEs, plotted against the actual boarding school data. Clearly, the model isn't well accounting for ramp up time in the data.