

# APPROXIMATION OF ESTIMATES IN THE SUSCEPTIBLE-INFECTIOUS-REMOVED EPIDEMIC MODEL

Changhyuck Oh

Department of Statistics, Yeungnam University, Gyeongsan, Republic of Korea

Received 2013-03-18; Revised 2013-03-18; Accepted 2013-03-21

## ABSTRACT

In this note, a simple approximation for the maximum likelihood estimates of infection and removal parameters used in the Susceptible-Infectious-Removed (SIR) epidemic model is presented. This approximation can be applied when the numbers of susceptible and infected individuals are observable only at discrete points in time. Since, in such cases, a closed form of the likelihood function is generally too complicated to obtain, the proposed approximation method represents an important advance. Simulation results show that the method yields approximations quite close to the maximum likelihood estimates obtained under continuous observation.

**Keywords:** Approximate Maximum Likelihood, Discrete Sampling Scheme, Infection and Removal Rate, Reproduction Number, Susceptible-Infectious-Removed Epidemic Model

## 1. INTRODUCTION

Over the last century, mathematical modeling of epidemiological phenomena has been used to understand, predict and control the spread of infectious diseases. Almost all mathematical models of diseases start from the same basic premise: that the population can be subdivided into a set of distinct classes, according to individual relations to the given disease. In the widely studied Susceptible-Infectious-Removed (SIR) model, individuals can be classified as (1) susceptible to a disease, (2) infected by it and (3) either removed from the population or rendered immune. When all instances of infection and removal are observed over a given time interval, the likelihood of SIR estimates for infection and removal can be maximized, as shown by Becker and Britton (1999); however, when observations occur only at discrete points in time, maximum likelihood is difficult to achieve, since no closed-form expression for the likelihood function can be obtained.

The SIR model can be regarded as a simple birth and death process in which being born is equivalent to becoming infected. For the birth and death process, a number of studies have focused on the problem of generating estimates based on observations at discrete time points. Under the equidistant discrete sampling

scheme in which two adjacent observation time points have the same distance, Kendall (1949) and Keiding (1974) explored the maximum likelihood estimation of the birth rate for the Yule process. Similarly, for a linear growth birth and death process, Keiding (1975) presented a maximum likelihood estimate for the so-called Malthusian parameter. Further, McNeil and Weiss (1977) provided diffusion approximation estimates of the sum of birth and death rates and the Malthusian parameter.

For the discrete sampling scheme (not necessarily using the equidistant constraint), several studies have investigated the approximation of maximum likelihood estimates for the infection rate of the simple stochastic epidemic model (Hill and Severo, 1969; Kryscio, 1972; Choi and Severo, 1988). Oh *et al.* (1991) presented approximations of the maximum likelihood estimate for the birth rate in a class of birth processes. Cronie and Yu (2010) studied several problems with maximum likelihood estimation in the immigration-death process. Chen and Hyrien (2011) considered quasi- and pseudo-likelihood estimation for a class of continuous-time multi-type Markov branching processes. Crawford *et al.* (2011) suggested estimation of parameters using the EM algorithm for a class of birth-death processes when a process is observed only at two time points (the beginning and the end of a given time interval) but did not consider the SIR model.

Under the SIR model, data augmentation techniques (Tanner and Wong, 1987) cannot be applied directly because of the difficulties in obtaining conditional expectations of the numbers of subjects in each of the three classes. To avoid such difficulties, Cauchemez and Ferguson (2008) approximated the SIR model with a diffusion process, but their approach assumed a large population size and would not be suitable for data collected in small communities or households. Following Oh *et al.* (1991), a simple method of approximating maximum likelihood estimates for the SIR model under continuous observation over a given time interval is presented. Unlike Cauchemez and Ferguson (2008), a large population size is not assumed, so our method should apply well to data from small communities.

This note is organized as follows. The SIR model is presented in Section 2, along with description of the methods of approximation. Section 3 considers numerical examples and provides simulation results. The final section includes a discussion and concluding remarks.

**1.1. Approximation**

Let  $X(t)$ ,  $Y(t)$  and  $R(t)$  denote, respectively, the number of susceptible, infected and immune or removed individuals at time  $t \geq 0$  of the SIR epidemic model with infection parameter  $\beta$  and removal parameter  $\gamma$  in a closed population of size  $N$ , that is, a Markovian process with infinitesimal transition probabilities:

$$P\{X(t+h) = x_h, Y(t+h) = y_h \mid X(t) = x, Y(t) = y\} = \begin{cases} (\beta/N)xyh + o(h), & (x_h, y_h) = (x-1, y+1), \\ \gamma y h + o(h), & (x_h, y_h) = (x, y-1), \\ 1 - [(\beta/N)xy + \gamma y]h + o(h), & (x_h, y_h) = (x, y), \\ o(h), & \text{otherwise} \end{cases}$$

for  $x = n, n-1, \dots, 0$  and  $y = 1, 2, \dots$  where  $X(t)+Y(t)+R(t) = N$ ,  $X(0) = n$ ,  $Y(0) = a > 0$ ,  $R(0) = 0$  and  $N = n+a$ . It is assumed that there are no latent periods and that once an individual is infected, (s) he becomes infectious. Only two of the random variables are independent. Consider the problem of estimating  $\beta$ ,  $\gamma$  and  $R_0 = \beta/\gamma$  when the process is observed only at a discrete set of time points,  $0 = t_0 < t_1 < \dots < t_k = T$ , where  $T$  is a fixed time point. The discrete sampling scheme gives data in the form:

$$D = \{(t_0, x_0, y_0) = (0, n, a), (t_1, x_1, y_1), \dots, (t_k, x_k, y_k) = (T, n-b, a+b-d)\} \tag{1}$$

where,  $x_i$  and  $y_i$  are the observed numbers of susceptible and infectious cases at time  $t_i$ , respectively and  $b$  and  $d$  are the number of infections and removals in  $(0, T]$ ,

respectively, with  $n-b \geq 0$  and  $a+b-d \geq 0$ . The corresponding likelihood function is the product of transition probabilities, which, in this case, are quite entangled; thus, it appears impossible to find explicit expressions for the maximum likelihood estimates of  $\beta$  and  $\gamma$  based on  $D$ .

Following Oh *et al.* (1991), simple approximations of the maximum likelihood estimates are presented that provide straightforwardly calculated initial values for an iterative procedure, yet achieve good accuracy and precision so long as the observation times  $t_1, \dots, t_n$  are sufficiently close together, such that  $(t_i, x_i, y_i)$  do not vary significantly between successive pairs.

To derive the approximations, it is first assumed that the process is observed continuously over the fixed time interval  $(0, T]$ , i.e., all infection and removal times that have occurred until  $T$  are observed. Furthermore, it is assumed that  $r$  infection or removal "events" have occurred and  $Z_j (j = 1, \dots, r+1)$  is the time between the  $(j-1)$ th and  $j$ th events; then,  $Z_1, \dots, Z_{r+1}$  are independent, exponentially distributed random variables with respective rate parameters. From this continuous sampling scheme, we have data of the form  $(Z_1, V_1) = (z_1, v_1), \dots, (Z_r, V_r) = (z_r, v_r)$  and  $Z_{r+1} > T - (z_1 + \dots + z_r) \equiv z_{r+1}$  if  $r \geq 1$  and  $z_1 > T$  if  $r = 0$ , where  $v_j = 1$  or  $0$  depending on whether infection or removal occurred at  $z_1 + \dots + z_j$  for  $j = 1, \dots, r$ , respectively. The corresponding log likelihood function becomes equal, when  $r \geq 1$ :

$$\log \ell(\beta, \gamma) \propto b \log \beta + d \log \gamma - \beta s_{XY} - \gamma s_Y \tag{2}$$

Where:

$$b = \sum_{j=1}^r v_j, d = r - b, s_{XY} = N^{-1} \sum_{j=1}^{r+1} x_{j-1}^c y_{j-1}^c z_j, s_Y = \sum_{j=1}^{r+1} y_{j-1}^c z_j$$

and for  $j = 0, 1, \dots, r$ ,  $x_j^c$  and  $y_j^c$  are the numbers of susceptibles and infections at time  $z_0 + \dots + z_j$ , respectively, with  $z_0 = 0$ . By differentiating Equation 2 with respect to  $\beta$  and  $\gamma$ , we can obtain the maximum likelihood estimates of  $\beta$  and  $\gamma$  under continuous observation:

$$\hat{\beta}_\infty = \frac{b}{s_{XY}}, \hat{\gamma}_\infty = \frac{d}{s_Y} \tag{3}$$

Corresponding to the results of Becker and Britton (1999),  $R_0$  can be estimated by:

$$\hat{R}_0 = \hat{\beta}_\infty / \hat{\gamma}_\infty \tag{4}$$

When the process is observed only at a discrete set of time points and the sample is in the form of  $D$ , the likelihood function is:

$$L(\beta, \gamma) = \prod_{i=0}^{k-1} l(\beta, \gamma | x_i, y_i, x_{i+1}, y_{i+1}) \tag{5}$$

Where:

$$l(\beta, \gamma | x_i, y_i, x_{i+1}, y_{i+1}) = P\{X(t_{i+1}) = x_{i+1}, Y(t_{i+1}) = y_{i+1} | X(t_i) = x_i, Y(t_i) = y_i\}$$

the probability generating function for which is given in Daley and Gani (1999). Note, however, that obtaining the maximum likelihood estimates of  $\beta$  and  $\gamma$  using Equation 5 seems infeasible. Under the discrete sampling scheme D, b and d are known but not their respective denominators  $s_{XY}$  and  $s_Y$  in Equation 3.  $s_{XY}$  and  $s_Y$  have been approximated using the method of Choi and Severo (1988) and Oh *et al.* (1991). In those studies, only one parameter (infection or birth rate) had to be estimated, whereas two parameters have to be estimated simultaneously here. For notational convenience,  $s_{XY}$  and  $s_Y$  are denoted by:

$$s_{XY} = \sum_{i=1}^k U_i \text{ and } s_Y = \sum_{i=1}^k V_i$$

Where:

$$U_i = N^{-1} \int_{t_{i-1}}^{t_i} x(t)y(t)dt \text{ and } V_i = \int_{t_{i-1}}^{t_i} y(t)dt$$

for  $i = 1, \dots, k$ . As  $U_i$  and  $V_i$  are not known under this sampling scheme, their trapezoidal approximations  $(t_i - t_{i-1})(x_i y_i + x_{i-1} y_{i-1}) / (2N)$  and  $(t_i - t_{i-1})(y_i + y_{i-1}) / 2$ , respectively, are employed. These approximations yield the estimators  $\hat{\beta}_k$  and  $\hat{\gamma}_k$ , as follows:

$$\hat{\beta}_k = \frac{2bN}{\sum_{i=1}^k (t_i - t_{i-1})(x_i y_i + x_{i-1} y_{i-1})} \tag{6}$$

$$\hat{\gamma}_k = \frac{2d}{\sum_{i=1}^k (t_i - t_{i-1})(y_i + y_{i-1})} \tag{7}$$

As is the case under continuous observation,  $R_0$  can be estimated by:

$$\hat{R}_{0k} = \hat{\beta}_k / \hat{\gamma}_k \tag{8}$$

The expressions  $\hat{\beta}_k$ ,  $\hat{\gamma}_k$  and  $\hat{R}_{0k}$  provide simple approximations for  $\hat{\beta}_\infty$ ,  $\hat{\gamma}_\infty$  and  $\hat{R}_{0\infty}$ , respectively. When  $k \rightarrow \infty$  in such a way that  $t_i - t_{i-1} \rightarrow 0$  for all  $1 \leq i \leq k$  so that  $\sum_{i=1}^k (t_i - t_{i-1})(y_i + y_{i-1}) \rightarrow \int_0^{t_k} y(t)dt$  and

$\sum_{i=1}^k (t_i - t_{i-1})(x_i y_i + x_{i-1} y_{i-1}) \rightarrow \int_0^{t_k} x(t)y(t)dt$ , we have  $\hat{\beta}_k \rightarrow \hat{\beta}_\infty \rightarrow \hat{\gamma}_k \rightarrow \hat{\gamma}_\infty$  and  $\hat{R}_{0k} \rightarrow \hat{R}_{0\infty}$ , respectively.

### 1.2. Numerical Examples and Monte Carlo Experiments

**Table 1** shows the behavior of approximations for the SIR process when  $\beta = 0.2$  and  $0.15$ ,  $\gamma = 0.1$ ,  $n = 180$ ,  $a = 20$  and  $T = 100$ . Ten simulation results are reported, five for  $\beta = 0.2$  and five for  $\beta = 0.15$ . For example, in the first realization generated using  $\beta = 0.2$ , there are 39 susceptibles and zero infected at the last observation time  $T_0 = 67.44$ , when no infected individuals remained. As a result, there are 141 infections and 161 removals from the beginning of the observation up to  $T_0$  and thus up to  $T$ . After  $T_0$ , when the epidemic is over, there are no more changes in the number of susceptibles and infected. In each realization, discrete observations are obtained under the equidistant sampling scheme with  $k = 14$ , so that  $t_i = i \times 7.14$  for  $i = 1, \dots, 14$ . The length of each interval is therefore approximately 7. Values of  $(x_1, y_1), \dots, (x_k, y_k)$  are then chosen to correspond to the given sampling scheme yielding data in the form of D given in Equation 1. For each simulation, we obtain values of  $\hat{\beta}_\infty, \hat{\gamma}_\infty$  and  $\hat{R}_{0\infty}$  using Equation 3 and 4 and  $\hat{\beta}_k, \hat{\gamma}_k$  and  $\hat{R}_{0k}$  using Equation 6, 7 and 8 with  $k = 14$ , as shown in **Table 1**. When  $\beta = 0.2$ , the relative percentage errors of  $\hat{\beta}_k$  when compared with  $\hat{\beta}_\infty$  are less than or equal to 2.6%, the relative percentage errors of  $\hat{\gamma}_k$  when compared with  $\hat{\gamma}_\infty$  are less than or equal to 3.0% and the relative percentage errors of  $\hat{R}_{0k}$  when compared with  $\hat{R}_{0\infty}$  are less than or equal to 2.2%. When  $\beta = 0.15$ , these values are 4.0, 4.8 and 0.8%, respectively. The preceding observation was corroborated by simulation of 10,000 replications of the SIR epidemic process for each combination of  $\beta = 0.2$  and  $0.15$ ,  $\gamma = 0.1$ ,  $T = 100$ ,  $a = 20$  and  $n = 180$ . The value for  $T$  is chosen to be close to  $2 \times (1/\gamma) \log(a+n)$ , two times the average duration for a major outbreak of the epidemic when  $n \rightarrow \infty$  (Bailey, 1975). Furthermore, when  $\hat{\beta}_k, \hat{\gamma}_k$  and  $\hat{R}_{0k}$  were computed for each realization in each combination, the means and standard deviations for the 10000 absolute deviations of the approximation for each of  $k = 14, 25, 50$  and  $\infty$  were also calculated, as shown in **Table 2**. For  $T = 100$ , the values of  $k = 14, 25, 50$  and  $\infty$  correspond to 7, 3.5, 2 and 1 days, respectively. These findings demonstrate the accuracy and precision of our approximations for the SIR process.

**Table 1.** Values of the last observation time  $T_0$ ,  $\hat{\beta}_\infty, \hat{\gamma}_\infty, \hat{R}_{0\infty}, \hat{\beta}_k, \hat{\gamma}_k, \hat{R}_{0k}$  and errors for 10 realizations of the SIR epidemic for  $\beta = 0.2$  and  $0.15$ ,  $\gamma = 0.1$ ,  $T = 100$ ,  $n = 180$ ,  $a = 20$ . In each realization, discrete observations are obtained under the equidistant sampling scheme with  $k = 14$ , about one week

$\beta$	$T_0$	$X(T_0)$	$Y(T_0)$	$\hat{\beta}_\infty$	$\hat{\gamma}_\infty$	$\hat{R}_{0\infty}$	$\hat{\beta}_k$	$\hat{\gamma}_k$	$\hat{R}_{0k}$	$e(\hat{\beta}_k)$	$e(\hat{\gamma}_k)$	$e(\hat{R}_{0\infty})$
.2	67.44	39	0	0.220	0.107	2.045	0.223	0.108	2.072	0.018	0.005	0.013
	100	26	1	0.204	0.100	2.032	0.200	0.098	2.048	0.016	0.024	0.008
	100	29	1	0.210	0.101	2.067	0.210	0.103	2.045	0.003	0.013	0.011
	88.37	32	0	0.197	0.097	2.045	0.192	0.094	2.052	0.026	0.030	0.004
	74.75	42	0	0.198	0.100	1.982	0.199	0.102	1.938	0.001	0.023	0.022
.15	81.62	66	0	0.145	0.097	1.488	0.150	0.102	1.476	0.040	0.048	0.008
	77.79	98	0	0.150	0.120	1.251	0.153	0.123	1.241	0.018	0.026	0.008
	100	52	2	0.153	0.093	1.643	0.158	0.095	1.663	0.030	0.018	0.012
	91.38	90	0	0.130	0.101	1.295	0.130	0.101	1.295	0.001	0.001	0.001
	100	107	1	0.109	0.095	1.155	0.111	0.095	1.163	0.012	0.005	0.007

$$e(\hat{\beta}_k) = |\hat{\beta}_\infty - \hat{\beta}_k| / \hat{\beta}_\infty, e(\hat{\gamma}_k) = |\hat{\gamma}_\infty - \hat{\gamma}_k| / \hat{\gamma}_\infty, e(\hat{R}_{0\infty}) = |\hat{R}_{0\infty} - \hat{R}_{0k}| / \hat{R}_{0\infty}$$

**Table 2.** Means and standard deviations in parenthesis of 10000 absolute deviations of  $\hat{\beta}_k, \hat{\gamma}_k$  and  $\hat{R}_{0k}$  from  $\hat{\beta}_\infty, \hat{\gamma}_\infty$  and  $\hat{R}_{0\infty}$ , respectively, for SIR with  $\beta = 0.2$  and  $0.15$ ,  $\gamma = 0.1$ ,  $T = 100$ ,  $n = 180$  and  $a = 20$  and  $k = 14, 25$  and  $50$  under the equidistant sampling scheme

$\beta$	$k$	$\hat{\beta}_k$	$\hat{\gamma}_k$	$\hat{R}_{0k}$
.20	14	0.004 (0.003)	0.002 (0.002)	0.018 (0.015)
	25	0.002 (0.002)	0.001 (0.001)	0.010 (0.008)
	50	0.001 (0.001)	0.001 (0.000)	0.005 (0.004)
.15	14	0.003 (0.002)	0.002 (0.002)	0.008 (0.007)
	25	0.002 (0.001)	0.001 (0.001)	0.005 (0.004)
	50	0.001 (0.001)	0.001 (0.001)	0.002 (0.002)

**Table 3.** Estimates of  $\beta, \gamma$  and  $R_0$  using smallpox data

$k$	$\hat{\beta}_k$	$\hat{\gamma}_k$	$\hat{R}_{0k}$
13	0.1642	0.1346	1.2199
22	0.1610	0.1324	1.2160
83( $\infty$ )	0.1633	0.1343	1.2159

**Table 3** shows the estimates of  $\beta, \gamma$  and  $R_0$  using a widely studied dataset from a smallpox epidemic in Abakaliki, Nigeria (Bailey, 1975). There are 83 daily observations, which, for practical purposes, can be reregarded as continuous. Based on this data, the discretely observed data for time intervals  $d = 7, 4$  and  $1$  can be obtained and the corresponding values for

$\hat{\beta}_k, \hat{\gamma}_k$  and  $\hat{R}_{0k}$  for  $k = 13, 22$  and  $83$ , respectively, can be evaluated. The length of the last time interval of observation varies according to the remainder of the 83 days recorded. Thus, for  $k = 13, 22$  and  $83$ , the lengths of the last time interval are 5, 2 and 1, respectively. Note that the estimates for  $\beta$  are quite close for all values of  $d = 7, 4$  and  $1$ , as are the estimates for  $\gamma$  and  $R_0$ .

## 2. CONCLUSION

In practical application, we must typically estimate the infection and removal rates for the SIR model under conditions of discrete-time observations. In this note, a very simple technique to approximate the maximum likelihood estimates of infection and removal rates and thus the reproduction number for continuously observed data, was proposed and used.

It is shown that as the number of time points of observation is increased, so that all adjacent times got closer, the approximations converged on the maximum likelihood estimates obtained from continuous observation. In the simulation results shown in **Table 2**, the means and standard deviations of the absolute deviations decreased as the number of observation time points increased. In simulations and for the Akabakili smallpox data, the estimates of Cauchemez and Ferguson (2008), which must assume a large population size, have not been compared to those of the proposed technique.

This technique might also apply to more complicated settings, including situations in which only the number of infections can be observed at each time interval and the number of initial susceptibles remains unknown.

### 3. ACKNOWLEDGEMENT

This work was supported by the 2012 Yeungnam University Research Grant.

### 4. REFERENCES

- Bailey, N.T.J., 1975. *The Mathematical Theory of Infectious Diseases and its Applications*. 2nd Edn., Griffin, London. ISBN-10: 0852642318, pp: 413.
- Becker, N.G. and T. Britton, 1999. Statistical studies of infectious disease incidence. *J. R. Statist. Soc.*, 61: 287-307. DOI: 10.1111/1467-9868.00177
- Cauchemez, S. and N.M. Ferguson, 2008. Likelihood-based estimation of continuous-time epidemic models from time-series data: Application to measles transmission in London. *J. R. Soc. Interface*, 5: 885-897. DOI: 10.1098/rsif.2007.1292
- Chen, R. and O. Hyrien, 2011. Quasi- and pseudo-maximum likelihood estimators for discretely observed continuous-time Markov branching processes. *J. Stat. Plan. Inference*, 141: 2209-2227. PMID: 21552356
- Choi, Y.J. and N.C. Severo, 1988. An approximation for the maximum likelihood estimator of the infection rate in the simple stochastic epidemic. *Biometrika*, 75: 392-94. DOI: 10.1093/biomet/75.2.392
- Crawford, F.W., V.N. Minin and M.A. Suchard, 2011. *Estimation for general birth-death processes*. Cornell University Library.
- Cronie, O. and J. Yu, 2010. *Maximum Likelihood Estimation in a Discretely Observed Immigration-Death Process*. 1st Edn., Department of Mathematical Sciences, Goteborg, pp: 21.
- Daley, D.J. and J.M. Gani, 1999. *Epidemic Modelling: An Introduction*. 1st Edn., Cambridge University Press, Cambridge, ISBN-10: 0521640792, pp: 213.
- Hill, R.T. and N.C. Severo, 1969. The simple stochastic epidemic for small populations with one or more initial infectives. *Biometrika*, 56: 183-96. DOI: 10.1093/biomet/56.1.183
- Keiding, N., 1974. Estimation in the birth process. *Biometrika*, 61: 71-80. DOI: 10.1093/biomet/61.1.71
- Keiding, N., 1975. Maximum likelihood estimation in the birth-and-death process. *Ann. Statist.*, 3: 363-72. DOI: 10.1214/aos/1176343062
- Kendall, D.G., 1949. Stochastic processes and population growth. *J. R. Statist. Soc.*, 11: 230-82.
- Krystio, R.J., 1972. On estimating the infection rate of the simple stochastic epidemic. *Biometrika*, 59: 213-214. DOI: 10.1093/biomet/59.1.213
- McNeil, D.R. and G.H. Weiss, 1977. A large population approach to estimation of parameters in Markov population models. *Biometrika*, 64: 553-558. DOI: 10.1093/biomet/64.3.553
- Oh, C., N.C. Severo and J. Slivka, 1991. Approximation to the maximum likelihood estimate in some pure birth process. *Biometrika*, 78: 295-299. DOI: 10.1093/biomet/78.2.295
- Tanner, M.A. and W.H. Wong, 1987. The calculation of posterior distributions by data augmentation. *J. Am. Statist. Assoc.*, 82: 528-540. DOI: 10.1080/01621459.1987.10478458