Mathematical Biosciences 256 (2014) 89–101

Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Estimating the basic reproduction number from surveillance data on past epidemics

Sorana Froda *, Hugues Leduc

Département de mathématiques, UQAM, C.P. 8888, succ. centre-ville, Montréal, Québec H3C 3P8, Canada

ARTICLE INFO

Article history:
Received 1 August 2013
Received in revised form 27 June 2014
Accepted 3 August 2014
Available online 26 August 2014

Keywords:
SIR model
Basic reproduction number
Non homogeneous death process
Non homogeneous Poisson process
Estimation
Surveillance data

ABSTRACT

In this paper, we consider the basic reproduction number, R_0, a parameter that characterizes the transmission potential of an epidemic, and explore a novel way for estimating it. We introduce a stochastic process which takes as starting points the classical SIR (susceptibles-infected-removed) models, deterministic and stochastic. The estimation method rests on an extremum property of the deterministic SIR model, and could be applied to past surveillance data on epidemic outbreaks, data gathered at different locations or in different years. Our estimators take into account some practical limitations, in particular the fact that data are collected at preassigned times. We derive asymptotic properties of the estimators and perform a simulation study to assess their small sample behavior. We illustrate the method on real data (from the USA Centers for Disease Control and Prevention site) and we point to various extensions to our approach, as well as practical implementation issues.

1. Preliminaries

1.1. Introduction

Most epidemic models (deterministic or stochastic) are of compartmental type. The SIR (susceptibles-infected-removed) type ones suppose the following: in a population free of disease (the *susceptibles*), a few *infected* (and infectious) primary cases are introduced; the primary cases infect other, secondary cases, and the epidemic spreads; in the end, most infected recover and become immune, while some die; both recovered and dead are *removed*, *i.e.*, they cannot be infected again. In the original SIR deterministic ODE (ordinary differential equations) model given in Kermack and Mc Kendrick [24, special case B], a "critical" number of susceptibles, the *relative removal rate*, is necessary in order to have an epidemic outbreak. (This is true in the general case as well, but in our paper we refer mainly to the simpler formula.) This condition comes to the fact that its counterpart, the *basic reproduction number*, R_0, see (2), must be greater than 1 [1]. In many other extensions of model (1), R_0 has retained its importance and significance as a threshold value [35,5]. It has been argued that an epidemic can be contained if the fraction of vaccinations is greater than $1 - 1/R_0$, whether R_0 is defined via an ODE model or not [39,2,7].

In this paper, we introduce a model where the course of the epidemic is essentially driven by the deterministic process, while incorporating the stochasticity inherent in field data. Indeed, it is generally agreed that, in large populations, deterministic systems can be used to describe the (approximate) behavior of the infectious process. Further, we focus on a new estimation method of R_0 which can resort to data provided by health agencies (surveillance data). These data are incomplete in the sense that only incidence (new cases) are monitored over time, since neither infected prevalence nor total removals, at some time t, are known. Moreover, the data are collected at discrete, preassigned times. Our easy to implement estimation method takes into account these practical limitations, and rests on a classical property of the deterministic SIR system. Our procedure has the advantage of proposing both point and interval estimators.

The presentation is structured as follows. We start from two classical SIR models and introduce a variant of these classical SIR models, as well as the idea behind our estimation method (Section 1.2). In Section 1.3, we compare the new approach to existing literature. Further, we describe two novel estimators and study their properties, exact and asymptotic (Section 2). Finally, we apply our method to simulated and surveillance data (Section 3), while possible extensions and a final conclusion are given in Section 4. Useful properties of the deterministic system are given in Appendix A, where we prove the main properties of the estimation method as well.

* Corresponding author. Tel.: +1 514 987 3000x3220.
E-mail addresses: froda.sorana@uqam.ca (S. Froda), hugues.leduc.14@gmail.com (H. Leduc).

http://dx.doi.org/10.1016/j.mbs.2014.08.007
0025-5564/

1.2. Two classical SIR models and a new model

The classical deterministic SIR (susceptibles-infected-removed) models presented in Kermack and Mc Kendrick [24] which are extensions of ideas in Ross and Hudson [37] assume that a population of fixed size, N, is divided into *susceptibles*, *infected* (and infectious), and *removed* (by immunization or death). Thus, they correspond to a situation where the infected eventually becomes immune (or dies), and where the population is, in large part, susceptible to the disease. In what follows we refer mainly to the special case (1) [24, formula (29)].

In the original development, $x(t)$ is the number of susceptibles, $y(t)$ the number of infected, and $z(t)$ the number of removed, $t \in [0,\infty]$, and satisfy the ODE system ($\beta_N > 0,\ \gamma_N > 0$)

$$\begin{cases} x'(t) = -\beta_N x(t) y(t), \\ y'(t) = \beta_N x(t) y(t) - \gamma_N y(t), \\ z'(t) = \gamma_N y(t), \end{cases} \tag{1}$$

of initial conditions $z(0) = 0$, $x(0) = x_0 > 0$ and $y(0) = y_0 > 0$, with $x_0 + y_0 = N$; usually $y(0)$ is negligible when compared with $x(0)$. The ratio $\rho_N = \gamma_N/\beta_N$, is the *relative removal rate* (see, e.g., [14, Chapter 2]). In general, the *basic reproduction number* R_0 is defined as follows: if one infected is introduced in a very large population of susceptibles only, *i.e.*, $x(0) \approx N$, the average number of secondary infections this infected produces during the period he is infectious is R_0. In model (1) this property can be translated into

$$R_0 = \frac{\beta_N N}{\gamma_N} = \frac{N}{\rho_N}, \tag{2}$$

because β_N is the contact parameter and $1/\gamma_N$ is the average time one is infectious (the full argument is given in Heesterbeek and Dietz [18] and is related to the *infectivity function*).

Further, set $\beta_N = \beta/N,\ \gamma_N = \gamma$ ($\beta > 0, \gamma > 0$). Then $\rho = \gamma/\beta$ is the relative removal rate per susceptible and the system satisfied by the proportions $(x(t), y(t), z(t))$, with initial value (x_0, y_0, z_0), is related to the system satisfied by the sizes $(\tilde{x}(t), \tilde{y}(t), \tilde{z}(t)) = (Nx(t), Ny(t), Nz(t))$ with initial value (Nx_0, Ny_0, Nz_0) in the following obvious way:

$$\begin{cases} \tilde{x}'(t) = -\frac{\beta}{N}\tilde{x}(t)\tilde{y}(t), \\ \tilde{y}'(t) = \frac{\beta}{N}\tilde{x}(t)\tilde{y}(t) - \gamma\tilde{y}(t), \\ \tilde{z}'(t) = \gamma\tilde{y}(t), \end{cases} \iff \begin{cases} x'(t) = -\beta x(t) y(t), \\ y'(t) = \beta x(t) y(t) - \gamma y(t), \\ z'(t) = \gamma y(t). \end{cases} \tag{3}$$

Both systems in (3) give the same value of R_0. Indeed, $\tilde{x}(t_Z) = \rho_N = \rho N$ and $R_0 = N/\rho_N = 1/\rho$. In what follows, wherever we relate population sizes to percentages we use the notation $(x(t), y(t), z(t))$ for percentages and $(\tilde{x}(t), \tilde{y}(t), \tilde{z}(t))$ for sizes.

It is well known in the ODE literature, that the behavior of the solution to (3) differs according to $x_0 > \rho$ or $x_0 < \rho$ and, in this last case, there is no epidemic, as $y(t)$ is strictly decreasing (see Appendix A). In the case where there is an epidemic, by considering the time t_Z where $z''(t) = y'(t) = 0$, one has from (3) the relation at the core of our estimation method,

$$\rho = x(t_Z) \iff \rho_N = \tilde{x}(t_Z); \quad R_0 = \frac{N}{\tilde{x}(t_Z)} = \frac{N}{Nx(t_Z)} = \frac{1}{x(t_Z)}, \tag{4}$$

which allows us to express R_0 as a function of $x(t_Z)$. Note that the second formula in (4) becomes $R_0 x(t_Z) = 1$, which is similar in spirit to various proposals due to Hethcote and his co-authors, as reviewed in Hethcote [20], or to the definition of R_0 given in Anderson and May [2] in terms of the fraction of the host population at equilibrium in endemic models. What is particular to our case and is exploited in the estimation is that all trajectories in the (x, y) phase plane pass through $(\rho, \max_t y(t))$, no matter the starting value (x_0, y_0) in $x_0 + y_0 \leqslant 1$, with $R_0 x_0 > 1$ (see [19]). Moreover, we propose an explicit model of random variation around the deterministic one which allows, among others, to compute confidence intervals.

In order to introduce the new approach we recall that the stochastic model which is the natural counterpart of the deterministic version (1), is given by a trivariate Markov chain in continuous time, $\{(X(t), Y(t), Z(t)),\ t \geqslant 0\}$ with state space $S \subset \mathbb{N}^3$ and initial values $(X(0), Y(0), Z(0))$. We will refer to it as the stochastic SIR model or Model 1 (see Appendix A, Section A.2). This type of model originates in the work of McKendrick [30]. The stochastic SIR model was studied by Bartlett [4], Kendall [23], as well as Whittle [41] and Ball [3]. In Model 1, the values $X(t), Y(t)$, and $Z(t)$ satisfy the condition $X(t) + Y(t) + Z(t) = N$, for all $t \geqslant 0$. Therefore, in the literature, it is customary to express Model 1 in terms of the bivariate process $(X(t), Y(t))$. Kurtz [25,26] proves that, if $N \to \infty$, the paths of the stochastic model of parameters $(\beta/N, \gamma)$ converge weakly to the solution of (3) with both stochastic and deterministic values translated into proportions.

The method developed in this paper stems from the following idea: although, conceptually, it makes sense to express $X(t)$ and $Z(t)$ in terms of $Y(t)$, as has been done for Model 1 [23], if we want to estimate the parameters in (1), or at least the ratio $\rho = \gamma/\beta$ and R_0, $Y(t)$ cannot be used, since it is not observed in current practice. Moreover, the cumulative number of removals $Z(t)$ is also unknown, but some studies report the cumulative number of deaths, which can be supposed to be a small proportion of the number of removals. What is actually measured is the incidence, *i.e.*, the new cases (newly infected), which corresponds to the decrease in the number of susceptibles in given time intervals. It is well-known that the expectations of the random variables in Model 1 do not satisfy (1) (see [22]).

By construction, in the stochastic SIR Model 1 pairs of random variables satisfy simultaneous surges and drops. Thus wherever $Y(t)$ increases by one unit, then $X(t)$ goes down by one unit while, wherever $Y(t)$ drops by one unit, then $Z(t)$ augments by one unit. Therefore, rather than modeling $Y(t)$, we propose to model the couple of "observed" susceptibles and removals $(X(t), Z(t))$.

Our interest is focused on the marginal processes $X(t)$ and $Z(t)$, and the bivariate process $(X(t), Z(t))$ is such that the expected value of the triplet $(X(t), N - X(t) - Z(t), Z(t))$ is the solution $(x(t), y(t), z(t))$ of the system (1), in the parametrization $(\beta_N, \gamma_N) = (\beta/N, \gamma)$ with $\beta > 0$, $\gamma > 0$. In particular, the expected values satisfy $x(t) + z(t) = N - y(t)$. For the marginal processes we obtain that:

a. $X(t)$ is a non homogeneous pure death process of rate $\mu_X(t) = \beta_N y(t) X(t)$.
b. $Z(t)$ is a non homogeneous Poisson (birth) process of rate $\lambda_Z(t) = \gamma y(t)$.

In other words, given that our estimation method is meant to be applied to larger populations, we assume that the course of the epidemic is essentially driven by the deterministic process, but we take into account the stochasticity inherent in field data. The bivariate process is given in Appendix A, Section A.2.

Finally, we introduce the process of the (cumulative) number of deceased observed at time $t \geqslant 0$. Let p be the fixed probability that a removal corresponds to a death and let V_j be an indicator variable such $V_j = 1$ if the jth removal is a death, and 0 otherwise. Thus $V_j,\ j = 1, 2, \ldots$ are i.i.d. Bernoulli variables with success probability p. We propose to model $\{D(t), t \geqslant 0\}$ as the compound non homogeneous Poisson process

$$D(t) = \begin{cases} \sum_{j=1}^{Z(t)} V_j, & Z(t) > 0, \\ 0, & \text{otherwise}. \end{cases} \tag{5}$$