10 Estimation in partially observed epidemics

In the previous chapter, statistical analysis was based on what we called complete observation of the epidemic process; both the times of infection and removal (recovery) for all infected individuals were observed. In real life such detailed data is rarely available. In the present chapter we study estimation procedures for less detailed partial data. The likelihood for partial data is usually cumbersome to work with, being a sum or integral over all complete data sets resulting in the observed partial data. Other estimation techniques can turn out to be much simpler. Below we present two general techniques which have been successful in several epidemic applications: martingale methods and the EM-algorithm.

The two techniques are applied to a specific estimation problem. First, we use martingale methods to derive an estimator for the basic reproduction number in the standard SIR epidemic, assuming that only the initial and final states are observed. Second, we consider a discrete time version of the models presented previously, introducing chain-binomial models, which are interesting in their own right. Assuming that the chains are observed we make use of the EM-algorithm to deduce an estimator of the transmission parameter. It is worth pointing out that pure likelihood methods can be applied in several cases, even when only partial data is available. If, for example, several small sub-populations (e.g. households) are observed, then ML-estimation is often possible from final size data. Estimation then consists of traditional ML-theory with the additional ingredient of potential numerical problems due to the complicated form of the likelihood (cf. the exact likelihood in Section 2.4). Addy et al. (1991) present estimation procedures for final size data of sub-populations, allowing for a heterogeneous community and also for transmission from outside the sub-population. The ideas extend the early work by Longini and Koopman (1982) who treat a homogeneous community and a special form of transmission dynamics.

10.1 Estimation based on martingale methods

The standard SIR epidemic $E_{n,m}(\lambda, I)$ was defined in Section 2.1. Suppose that we have collected data from such an epidemic and that data consist of knowing how many individuals have been infected during the course of the epidemic, besides knowing the initial state (i.e. the number of initially susceptible and infectious individuals respectively). In many cases the initial state may not be known. The initial number of infectives is usually small thus not causing much uncertainty; the substantial problem lies in knowing how many individuals are initially susceptible to the disease. Quite often some individuals are immune even before the disease is introduced, perhaps due to an earlier outbreak of the same or a similar disease. Then the number of susceptibles has to be estimated prior to the outbreak, using traditional statistical methods. Here we assume this number to be known thus neglecting such uncertainty.
Our data consists of $n = X(0)$, $\mu n = Y(0)$ (the initial state) and the final number infected $Z = n - X(T)$, and implicitly knowing that $Y(T) = 0$ since there are no infectious individuals at the end of the epidemic. As observed in Section 2.4 the exact distribution of the final size of the epidemic has a numerically complicated form, thus not enabling much help in estimation when $n$ is large. One way to estimate is of course to use the central limit theorem stating that $Z$ is approximately normally distributed and equating the mean to the observed value $Z$. Here we take a different approach based on martingale methods, a method applicable in many other cases.

From Chapter 9 we know that

$$
\int_0^u f(t) \left( dN_1(t) - \lambda_0 \bar{X}(t) Y(t) dt \right) - \int_0^u g(t) \left( dN_2(t) - \gamma_0 Y(t) dt \right)
$$

is a martingale for any choice of predictable (left-continuous) processes $f = \{ f(t); t \geq 0 \}$ and $g = \{ g(t); t \geq 0 \}$. The ML-estimator $\hat{\lambda}$ was obtained by equating the martingale to 0 for the choice $f(t) = 1/\lambda$ and $g(t) = 0$, and $\hat{\gamma}$ from $f(t) = 0$ and $g(t) = 1/\gamma$. This method can no longer be used in the case of final size data: for example $\int_0^u \bar{X}(t) Y(t) dt$ appearing in $\hat{\lambda}$ is not observed. However, the same idea of equating a martingale to its mean, a special form of the method of moments, can be adopted by choosing $f$ and $g$ cleverly so that the resulting martingale only depends on observed quantities. We want to choose $f$ and $g$ such that the ‘$dt$’ terms cancel out. This happens if $f(t) = \gamma_0 / (\lambda_0 \bar{X}(T-))$ (note that $f(t)$ is left-continuous) and $g(t) = 1$. Let $M^{(n)}$ denote the resulting zero-mean martingale divided by $\sqrt{n}$ and let $\theta_0 = \lambda_0 / \gamma_0$ be the basic reproduction number. That is,

$$
\sqrt{n} M^{(n)}(u) = \int_0^u \frac{1}{\theta_0 \bar{X}(t-)} \left( dN_1(t) - \lambda_0 \bar{X}(t) Y(t) dt \right) - \int_0^u (dN_2(t) - \gamma_0 Y(t) dt)
$$

$$
= \frac{1}{\theta_0} \int_0^u \frac{1}{\bar{X}(t-)} dN_1(t) - \int_0^u dN_2(t)
$$

$$
= \frac{n}{\theta_0} \left( \frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n-(N_1(u)-1)} \right) - N_2(u).
\quad (10.1)
$$

The last equality is true since $X(t) = n - N_1(t)$. For $u = T$, the end of the epidemic, its value is determined by the final size data since $N_1(T) = Z$ and $N_2(T) = \mu n + Z$. Solving the equation $M^{(n)}(T) = 0$ leads to the estimator

$$
\hat{\theta} = \frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n-(Z-1)} \quad \frac{n}{Z + \mu}.
\quad (10.2)
$$

To see that the estimator is reasonable, we look at its large population approximation using the approximation $1/n+1/(n-1)+\cdots+1/(n-(Z-1)) \approx -\log(1-Z)$ implying that

$$
\hat{\theta} \approx -\log(1-Z)/(Z + \mu).
$$