**Practical: Fit an epidemic**

Scenario: Country Q is a small island country with a population size of 100,000. In early 2019, an infectious disease outbreak was observed in the Country Q resulting a large number of infections. Country Q has never experienced this infectious disease before, i.e. everyone was susceptible to this disease at the beginning of this outbreak. It is known that the outbreak was due to a group of tourists who were infectious during their visit. The government also followed up close contacts of these infected tourists. We have been able to construct a database of epidemiologic data of this outbreak in Country Q.

Database: The database contains two csv files. The file “daily\_incidence\_observed.csv” records the daily number of new infections observed by government, where day = 1 corresponds to 3 January 2019.

The file “data\_interval\_between\_symptom\_onset.csv” contains two columns: the first column is the ID of the infectious tourist and the second column is the time interval (days) between the dates of symptoms onset of the infector-infectee pairs.

Answer the questions below:

Q1: Which of the following epidemiologic parameter could be estimated from the data in the file “data\_interval\_between\_symptom\_onset.csv”, incubation period, duration of infectiousness, serial interval or generation time? Why? [10 marks]

Q2: Using AIC, determine which of the following parametric models is most suitable for estimating the distribution of the epidemiologic parameter in Q1: exponential, binomial, or lognormal? Justify your answer and quote the parameter estimate(s) for the model you select. What is the mean of the epidemiologic parameter in Q1 based on the model you selected? [15 marks]

Note for Q2:

The Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models, which provides a means for model selection. Suppose that we have a statistical model of some data. Let *L* be the maximum value of the likelihood function for the model; let *k* be the number of estimated parameters in the model. Then the AIC value of the model is



Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Hence AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters.

Q3: Simulate a deterministic SIR model for the Country Q. We know that Country Q has a population size of 100,000 and there were 90 infectious individuals on 3 Jan 2019. The disease surveillance department estimated that reporting rate was about 15%, i.e. 15% of the infections had been reported. Assume that the mean duration of infectiousness in the SIR model is the same as the mean of the epidemiologic parameter in Q1. How can you estimate the basic reproductive number *R*0? If so, what’s your best estimate of *R*0? [75 marks]

Note for Q3:

First, you may try different values of *R*0 and see if the resulting epidemic curve looks like the epidemic curve from “daily\_incidence\_observed.csv” by just eyeballing.

To measure how “close” the epidemic curve generated by a certain *R*0 is to the observed epidemic curve, we could make use of the maximum likelihood method. Suppose that the observed incidence *Yt* on day *t* is a Poisson random sample with expectation of the product of true incidence *Ct*  and the reporting rate *p*:



The probability of observing *Yt* if the true incidence on day *t* is *Ct* (we can use the Poisson probability mass function “dpois” in R)*:*



Under a particular *R*0 value, from the SIR model we could have a series of “true” incidence *Ct* (*t* = 1, 2, 3, …) and the probability of observing the incidence series of *Yt*is:



We want to look for a particular *R*0 under which it is most likely that we can observe the given epidemic curve. Therefore, our best estimate of *R*0 is the one which could maximize the above probability (i.e. the likelihood function).

To find out the best estimate of *R*0, we could calculate the above likelihood under different *R*0 values and see which *R*0 value has the maximum likelihood.

Here is a useful note on how to estimate model parameters by maximum likelihood: <http://daphnia.ecology.uga.edu/drakelab/wp-content/uploads/2014/07/likelihood.pdf>