MAM5220 Statistical Techniques for Computational Biology
MA35210 Topics in Biological Statistics

Epidemiology Workbook

You will be assessed on your answers to questions 2 to 5

Introduction

A well-established model for outbreaks of infectious diseases (epidemics) is the so-called SIR model. Here the population (of size N) is split into three mutually exclusive subsets:

S = susceptible (people who can contract the disease)

I = infectious (people who have and can transmit the disease)

R = removed (people who've had the disease and are now immune)

The total population is initially assumed to be fixed, i.e. N=S+I+R

In order to build the model of differential equations, we need two parameters for the basic model, an infection rate, β, and a recovery rate, γ.

The basic model, which includes terms for the rates of change of the three compartments (S,I and R), is as follows:

$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dI}{dt} = \beta SI - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

(These can be read as "the rate of change of susceptibles with time = -βSI", etc.)

Question 1

This question introduces the idea of using the Euler method to give an approximate solution to a system of (ordinary) differential equations.

The basic idea of the Euler method is to approximate the continuous time derivative (e.g. dS/dt) with a discrete time version, where we take small increments of time.

For the susceptibles compartment, we write

$$\frac{dS}{dt} \approx \frac{\Delta S}{\Delta t} = \frac{S(t+\Delta t) - S(t)}{\Delta t},$$

where ΔS (read "delta S") is a small change in the number of susceptibles and Δt ("delta t") is a small time increment. The top part (the numerator) of the right hand side of the equation corresponds to the number of susceptibles at time t+Δt minus the number of susceptibles at time t, i.e. the change in the number of susceptibles in the small time increment Δt.

We rearrange to obtain

$$S(t+\Delta t) \approx S(t) + \Delta t \frac{dS}{dt}$$

$$= S(t) + \Delta t(-\beta S(t) I(t))$$

$$= S(t) - \beta S(t) I(t) \Delta t,$$

which expresses the number of susceptibles after a small time increment, Δt, in terms of things we know at time t (i.e. S(t) and I(t)).

Similarly, for I(t+Δt) we obtain

$$I(t+\Delta t) \approx I(t) + \Delta t \frac{dI}{dt}$$

$$= I(t) + \Delta t(\beta S(t) I(t) - \gamma I(t))$$

$$= I(t) + \beta S(t) I(t) \Delta t - \gamma I(t) \Delta t.$$

Here is a copy of the R script from the lecture notes (slide headed "Sample code for Euler's method") to implement this differential equation model.

```
# Example code for Euler's method
S0<-9000                        # Starting value for
susceptibles
I0<-50                          # Starting value for
infectious
N<-10000                        # Total population size
beta<-0.0001                    # infection rate
gamma<-1/50                     # recovery date
delta.t<-0.1                    # Small time increment
N.time.steps<-3000              # Number of time steps
S<-numeric(N.time.steps+1)      # Blank vector to receive S
values
I<-numeric(N.time.steps+1)      # Blank vector to receive I
values
S[1]<-S0                        # Initial value for S
I[1]<-I0                        # Initial value for I
for (i in 1:N.time.steps){      # Iterative loop to define
the next value of S and I
     S[i+1]<-S[i]-beta*S[i]*I[i]*delta.t
     I[i+1]<-I[i]+beta*S[i]*I[i]*delta.t-
gamma*I[i]*delta.t
}
time.vector<-seq(0,N.time.steps*delta.t,by=delta.t)    #
The time points
plot(time.vector,I,type="l",xlab="Time",ylab="I")

                                # Plotting the results
```

Note we are using Δt=0.1, so 3000 time steps corresponds to time=300 at the end of the simulation.

a) Modify the above code to find the number of susceptibles and infectious (which will not necessarily be whole numbers), after 10 time units if we start off with 5000 susceptibles, 100 infectious and 900 removed (i.e. N=6000), using the following parameter values

$$\beta = 0.00008$$

$$\gamma = 0.02$$

b) Produce side by side plots, with suitable labeling, of the number of susceptibles against time and the number of infectious against t over the time interval [0,200]

Question 2

In this question we consider the problem of fitting parameters to a simulated epidemiology data set.

Suppose we have a city of population one million (N=1000000). We are given data about the number of infectious people recorded the same day each week over a period of 30 weeks for an epidemic of a disease. We will use a time increment $\Delta t=0.1$ as previously. (We will use an SIR model without including the birth and death rate, i.e. the same model as Question 1.)

In the table below the number of people infected on a particular day each week are shown. The time index increases by 70 each week because there are 70 increments of 0.1 day in a week. The data are available in the .csv file epidemiology.data.csv on the Blackboard page for the Epidemiology section of MA35210 or MAM5220.

Week	Time index	No. infected	Week		No. infected
0	1	61	16	1121	10220
1	71	195	17	1191	8542
2	141	499	18	1261	3896
3	211	1668	19	1331	5297
4	281	3731	20	1401	1985
5	351	10104	21	1471	1761
6	421	19276	22	1541	1044
7	491	55190	23	1611	775
8	561	59881	24	1681	342
9	631	82025	25	1751	253
10	701	86428	26	1821	155
11	771	69840	27	1891	132
12	841	48055	28	1961	59
13	911	42489	29	2031	58
14	981	25511	30	2101	35
15	1051	16694			

a) Create a plot of number infected against week index with appropriate titles and axis labels. Describe the epidemic curve.

(6 marks)

We write $S_0=\theta N$, where θ is a proportion (i.e. $0 \leq \theta \leq 1$).

b) Modify the code of Question 1, specifying $R_0=10$ and mean infectious period two weeks ($\gamma=1/14$), to produce the graphs of number of infectious against time predicted by the model for $\theta = 0.25, 0.5, 0.75, 1$.

(4 marks)

A commonly used technique for measuring how well a particular choice of parameters fits the data is a sum of squares. Here is the R code for a function called errorSS. The argument epidemiology.data is the data frame you have loaded in (epidemiology.data.csv).

```
errorSS<-
 function(epidemiology.data,N=1000000,R0,D,theta,N.time.steps=2
 100,delta.t=0.1){
 gamma<-1/D
 beta<-R0*gamma/N
 S0<-theta*N
 I0<-epidemiology.data$No.infected[1]
 sample.index<-epidemiology.data$Time.index
 S<-numeric(N.time.steps+1)
I<-numeric(N.time.steps+1)
S[1]<-S0 # Initial conditions
I[1]<-I0 # Initial conditions
for (i in 1:N.time.steps){
 S[i+1]<-S[i]-beta*S[i]*I[i]*delta.t
 I[i+1]<-I[i]+beta*S[i]*I[i]*delta.t-gamma*I[i]*delta.t
}
errorSS<-sum((epidemiology.data$No.infected-
 I[sample.index])^2)
return(errorSS)
}
```

Suppose we first consider the situation where we consider the mean infectious period fixed (D=14 days). We fix $\theta = 0.3$. If we were not told the R_0 value for this epidemic, we may wish to try to infer it from the data.

c) Calculate the $\log_e$ (error sum of squares) for R_0 taking values in the vector $(6, 6.2, 6.4, \ldots, 13.6, 13.8, 14.0)$ and present your results visually, commenting on the form of the plot. Explain why taking logarithms is valid in this case.

(6 marks)

Suppose we wish now to infer **both** R_0 and D (the mean infectious period) from the data. Allow R_0 to vary over the same vector of values as in part c), and allow D to vary over the values in the vector $(10, 10.2, 10.4, \ldots, 17.6, 17.8, 18)$. (In order to store the error sums of squares obtained you will need a matrix.)

d) Report the minimum $\log_e$ (error sum of squares) obtained along with the corresponding R_0 and D values.

(4 marks)

e) Use the `persp()` function to visualize the surface of $\log_e$ (error sum of squares) over the ranges of R_0 and D provided. (The command `persp(x,y,z,theta=0,phi=15)` will

produce a 3D-surface plot of the matrix of values `z` over the ranges of `x` and `y`. The arguments `theta` and `phi` enable you to change the viewing angle – experiment with these to find an appropriate view. See the help page for `persp` for more details.) Comment on what you observe using the `persp` function here.

(8 marks)

f) For the optimal values of R_0 and D found in part d) (and θ=0.3), produce a plot with the fitted model curve (i.e. no. of infected against time in weeks) and the data points on the same graph. (HINT: plot the data first and then add the points from the model using the `points()` function. If you use (`time.vector-A)/B` as the x-axis for suitably chosen A and B, then this will correspond to time in weeks.) Comment on what you observe.

(6 marks)

(Question = 34 marks)

Question 3

The basic reproduction number, R_0, specific to a particular infectious disease, is given by

$$R_0 = \frac{\beta N}{\gamma},$$

and its interpretation is the number of secondary infections caused by a single infective introduced into a population made up entirely of susceptible individuals (S_0=N).

Given the following parameters

$$N = 5000000$$

$$R_0 = 20$$

$$\gamma = 0.1$$

$$\Delta t = 0.1$$

a) Calculate β, the infection rate, using the R_0 value

(2 marks)

b) Let $S_0 = \theta N$. Investigate the effects of altering θ and I_0 on the time course of infectious against time. Choose a range of values for θ in the interval (0,1] and consider a range of possible values of I_0, including $I_0 = 1$. (It is sensible to vary θ and I_0 independently, i.e. holding the other fixed.) You may wish to plot multiple graphs on the same axes to save space and to help with visualization.

(20 marks)
(Question = 22 marks)

Question 4

Here we extend the model to include births and deaths. We assume a constant birth rate, B births per unit time, all of whom enter the S pool (the susceptibles). We also assume a constant per capita mortality rate, μ. The equations become

$$\frac{dS}{dt} = B - \beta SI - \mu S$$

$$\frac{dI}{dt} = \beta SI - \gamma I - \mu I$$

a) What do the three parts of the right hand side of the first equation represent?

(3 marks)

For the next part of this question, please make use of the following function.

```
generate.S.I.by.time.vital.dynamics<-
 function(N.time.steps,delta.t=0.1,S0,I0,R0,gamma,mu,N){
 S<-numeric(N.time.steps+1)
 I<-numeric(N.time.steps+1)
 S[1]<-S0
 I[1]<-I0
 beta<-R0*gamma/N
 for (i in 1:N.time.steps){
    S[i+1]<-S[i]+mu*N*delta.t-beta*S[i]*I[i]*delta.t-
 mu*S[i]*delta.t
    I[i+1]<-I[i]+beta*S[i]*I[i]*delta.t-gamma*I[i]*delta.t-
 mu*I[i]*delta.t
 }
 time.vector<-seq(0,N.time.steps*delta.t,by=delta.t)
 out<-list(S=S,I=I,time.vector=time.vector)
 return(out)
}
```

b) The basic reproduction number for diphtheria may be taken as R_0=6. Suppose we wish to model a diphtheria epidemic in a city of population 800000. Suppose we know that γ=0.1 and the birth rate is 30 per 1000 population per year (so the daily rate, B=30×(800000/1000)/365=65.75). (We assume that the birth rate=overall death rate, i.e. $B = \mu N$.) Using Δt=0.1 as before, an initial number of infected equal to 50 and an initial number of susceptibles, $S_0 = 0.02N$, produce a plot showing how the number of infectious people varies over time. (Make sure the time axis extends far enough to capture multiple peaks in the time course.) Comment on the form of the plot.

(5 marks)

c) Investigate the effect of changing R_0 over the values [6, 8, 10, 12]. Clearly describe the effect on the graph of number of infectious over time.

(6 marks)

d) We now consider the effect of changing the birth rate, B. We fix R_0 at 10. Firstly allow B to vary over the set {100,200,300,600} (where these numbers represent daily birth rates). Produce four graphs of infectious against time and comment on the impact of increasing B on the form of the graphs.

(4 marks)

e) Now, with R_0 still fixed at 10, allow B to vary over the set {20,30,40,…,130,140}, i.e. 20 to 140 in increments of 10. By setting the run time of the simulation to a suitably large amount, calculate, for each value of B, the following quantities:
    - An approximation of the equilibrium number of infectious people.
    - The heights of the first four peaks.
    - The times of the first four peaks.

Find a way to represent these three measurements in a table and graphically, with B on the x-axis. (Produce three graphs, one for each bullet point. For each of the second and third bullet points, the graphs should have four lines, maybe with four colours, with one line for each of the four peaks.) Report your findings.

[Hint: If the run time is large, the fluctuations should eventually become very small. If you average over, say the final 10000 time increments, once the fluctuations are small, that should give you the first bullet point. For the others, use the fact that the differences between consecutive points in the series are positive on the way up to a maximum and negative on the way down from a maximum. If you use the `diff()` function (which calculates differences between consecutive points) and the `sign()` function, which returns +1 or -1, you should be able to calculate the times and heights of the first four peaks. (You'll need to difference twice and think of the appropriate condition with signs.)]

(20 marks)

(Question = 38 marks)

Question 5

In lectures we saw an example of extending the SIR model with births and deaths to incorporate seasonal forcing. The model in this case is

$$\frac{dS}{dt} = B - \beta(t)SI - \mu S$$

$$\frac{dI}{dt} = \beta(t)SI - \gamma I - \mu I,$$

where $\beta(t) = \beta_0(1 + \alpha\cos(2\pi t))$, and t is measured **in years**.

Modify the function

`generate.S.I.by.time.vital.dynamics`

given in Question 4 to create a function

`generate.S.I.by.time.vital.dynamics.seasonal.forcing.`

Suppose we have the following parameters:

$$N = 5000000$$

$$\gamma = \frac{1}{5}$$

$$R_0 = 14$$

$$S_0 = 0.04N$$

$$I_0 = 100 \times 5/30$$

$$\mu = \frac{100000}{365N}$$

Use your function to plot the curves of number of infectious against time for time from 0 to 20 years for α taking values 0, 0.25, 0.5 and 0.75. Include your function in your answer.

(16 marks)

(Total = 110 marks)