**MTBI 2015 Homework #2**

**Please submit with name format Towers\_hwk2\_<first name>\_<last name> Please provide your R and C++ files, and a Word file that gives the output to your screen, plots, etc.**

**All code must conform to good coding practices, as described in** [**http://sherrytowers.com/2012/12/14/good-programming-practices-in-any-language/**](http://sherrytowers.com/2012/12/14/good-programming-practices-in-any-language/) **and all plots must conform to good plotting practices, as described in** [**http://sherrytowers.com/2013/01/04/good-practices-in-producing-plots/**](http://sherrytowers.com/2013/01/04/good-practices-in-producing-plots/)

**Question 1**

You will be given copies of four prospectus documents written by your fellow students.

Read each prospectus, and fill in the score sheet below with the rubric provided.

Scoring rubric:

* **Motivation:** how well did the student motivate the project proposal. Is it clear why you should care about this topic? Were citations and/or links to relevant literature provided? Score 1=poorly motivated to 5=well motivated.
* **Objective:** was the proposed model appropriate to the motivation? Did the student give the equations of the model? Did the model make sense? Did the student provide references related to the model? Is there any indication given as to whether or not the model is novel? Score 1=inappropriate or poorly described model to 5=appropriate and well-described novel model (4 would be an appropriate and well-described model that is not necessarily novel).
* **Description of sources of data:** are the sources of data well described? Can you use the citations and links provided to actually find the data the student mentions? (note: look the sources of data up for each prospectus! If you state that the data are well described, but I found that I couldn’t find them based on the information provided, you will get a poor grade on this part). Score 1=data not described to 5=data very well described and found to be available
* **Appropriateness of the data for fitting the model:** Was the proposed data appropriate for the model in that it described the time evolution of at least one compartment in the model? Note that the models we have been considering all involve rates of flow from one compartment to another (ie; variation in time). Thus, fitting the models requires time series data; data that have repeated measurements in time. One thing you need to watch out for is whether or not the data described are in fact time series data, or just data at one time point. Data at one time point cannot, by themselves, be used to fit the parameters of the models we examine. Score 1=data do not appear to be appropriate to model and/or cannot be used for fitting because the data are not time series measurements, to 5=data appear to be appropriate to the model and are time series measurements of some quantity related to the model.

Score Sheet: rubric scores between 1(poorly done) to 5(well done). For each project, complete the following score sheet on a separate page.

Project Name:

Scores: 1=(poorly done) to 5=(well done)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Motivation | Description  Of Data | Objective | Appropriateness  Of Data | Total  Score |
|  |  |  |  |  |

Comments on how prospectus could be improved:

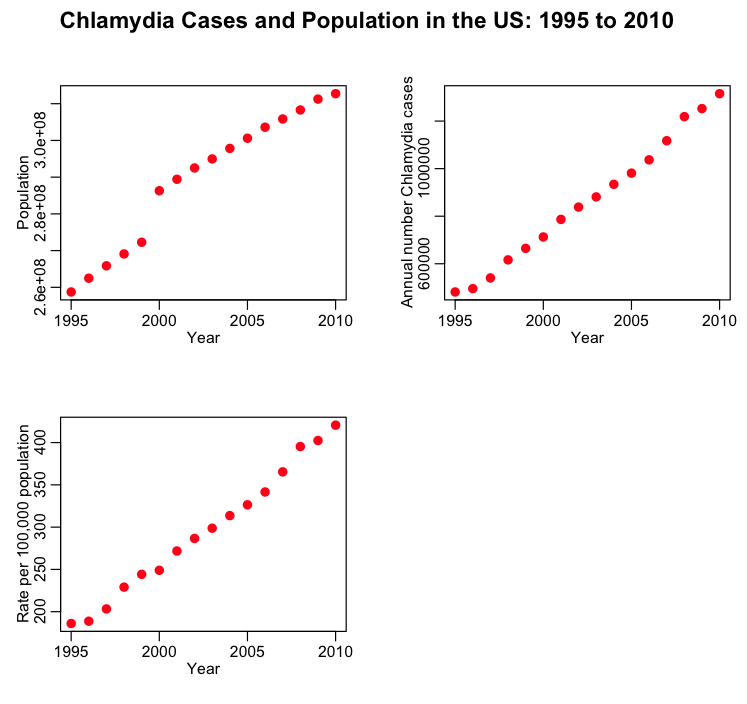
**Question 2**

From the CDC WONDER website, I have downloaded the STD incidence data from 1984 to 2013: <http://wonder.cdc.gov/std.html>

I’ve put this data in the file: <http://www.sherrytowers.com/cdc_std_data.txt>

In this exercise, we will examine the US Chlamydia incidence data between 1995 to 2010.

a) Write the R code to read in the file, and select the Chlamydia data in the appropriate time range. Make the following plot of the number of cases and population size by year:



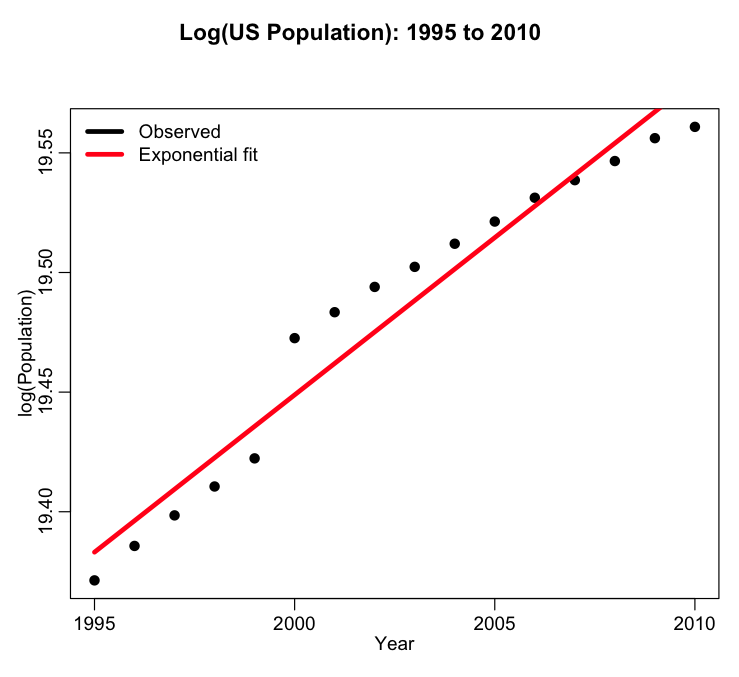
b) Obviously, from the figure, the US population is growing. In R, use the lm() linear regression model to regress log(Population) on the Year:

myfit = lm(log(Population)~Year)

summary(myfit)

rate\_of\_rise = myfit$coef[2]

The exponential rate of rise in the population is the birth rate, lambda, minus the death rate, mu. In the US, people live approximately 80 years. Thus, from your fit, what is the inferred average birth rate, lambda?



c) You do not gain immunity from Chlamydia upon treatment and recovery.

Using methods in the R deSolve library, code up the equations for the following SIS model

dS/dt = -beta\*S\*I/N + lambda\*(S+I) – mu\*S + alpha\*I

dI/dt = +beta\*S\*I/N – mu\*I – alpha\*I

where N = S + I, beta is the annual contact rate times probability of transmission upon contact, and alpha is the rate at which infected people are treated and return to the susceptible class.

Note that Chlamydia does not go away on its own, thus alpha\*I represents the approximate number of new people getting treatment each year (ie; the incidence… the observed number of newly identified cases).

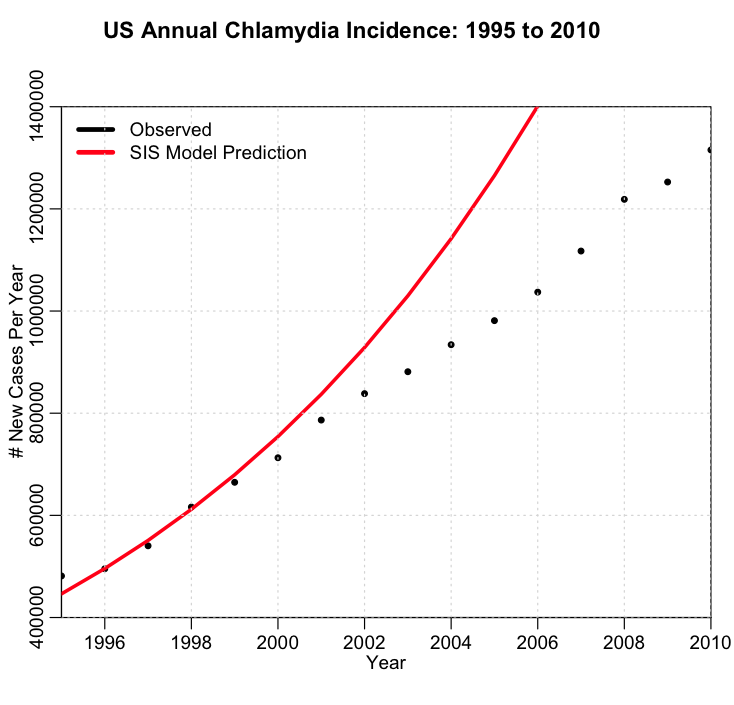
d) What are the underlying assumptions of this model:

* What does it assume about how the population mixes? Do you think this is a valid assumption?
* What does it assume about the probability distribution for the sojourn in the infected class before returning to the susceptible class? Do you think its realistic?

How would you change this model to make it more realistic?

e) Assume that the initial prevalence I/N in 1995 is I/N=f\_1995, and that the initial population at that time is the observed US population in 1995 from the data file.

To check whether or not you have coded up the model correctly, if you set the parameters to f\_1995 = 0.0025, beta=0.81, and alpha = 0.69, you should get the following plot:



f) Write the code to randomly sample hypotheses for beta, alpha, and the initial fraction of prevalence in the population in 1995, f\_1995. Assume that 1/mu=80 years, and that lambda is what you infer from the linear regression study. Set the initial population size to the population in 1995.

For each randomly sampled set of parameter hypotheses, get the SIS model estimate of the expected incidence per year (note that this is **not** the prevalence… the number of newly identified cases per year is the number of newly treated cases per year; alpha\*I).

The CDC study found that Chlamydia is under-diagnosed in the US, likely because most infections are asymptomatic: <http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6338a3.htm>

They estimate that around 2010 1.7% of the population between the ages of 14 to 39 was infected with Chlamydia (this is the estimated prevalence, I, at that time). In this analysis, we will make the simplifying assumption that the only people who get Chlamydia are people 14 to 39, which compose 35% of the US population in 2010. Thus the overall Chlamydia prevalence in 2010 was approximately I/N=f\_2010=0.017\*0.35=0.006.

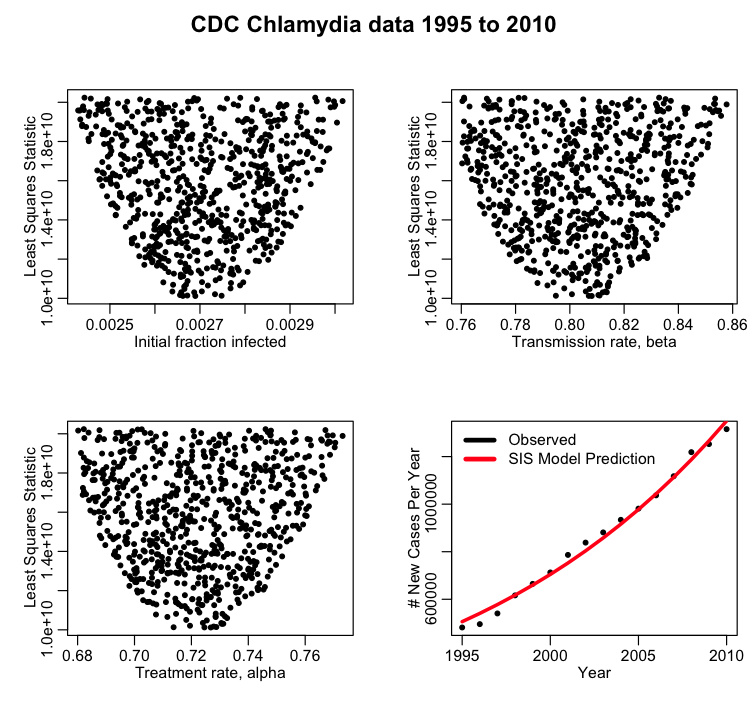
Write the code to calculate the Least Squares statistic comparing the model to the data. Because we know what the prevalence was in 2010, add a term onto the Least Squares statistic like this

LS = sum((data-model)^2)

LS = LS + (f\_2010\*N\_2010-I\_2010)^2

Where I\_2010 is your model estimate for I in 2010, and N\_2010 = S\_2010 + I\_2010

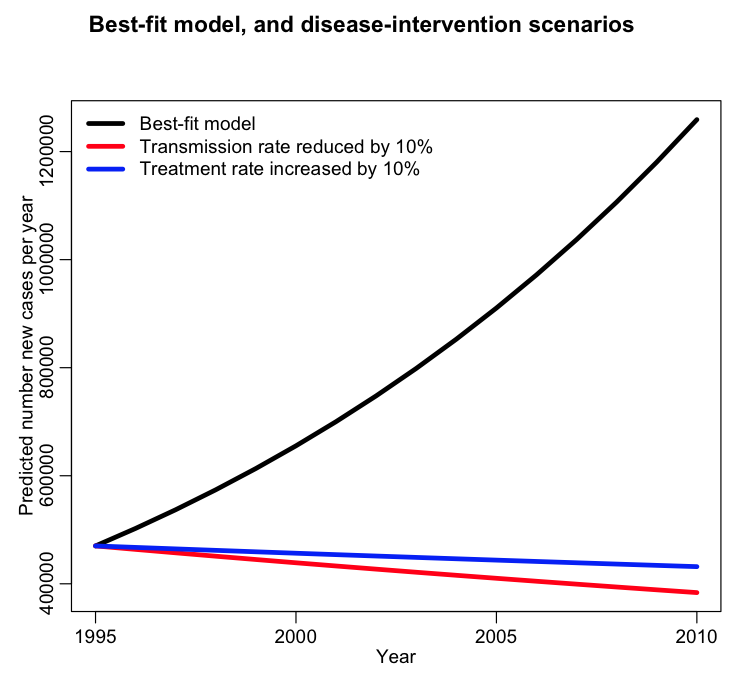
Now randomly sample the initial fraction infected, alpha, and beta in appropriate ranges, and calculate this Least Squares statistic. Repeat the process at least 100,000 times, to produce the following plot (yours won’t be exactly the same because your random seed will be different than mine).



g) What is the best estimate for the average time between someone catching Chlamydia, and getting treatment? Does this estimate seem plausible?

h) What is the best estimate for the average number of contacts, sufficient to transmit infection, that people in the population make each year? Does this seem plausible? Given that we modelled the entire population of the US, rather than the at-risk groups (ie; the sub-groups that have sex), what can you infer about what the true average number of contacts that transmit infection is.

i) When fitting model parameters to data, it is often desirable to then use the fitted model to examine control scenarios. Produce the following plot that examines two scenarios compared to the best-fit model; in the first reduce the transmission rate by 10% (which could, for instance, simulate increased condom use in the population). In the second, increase the treatment rate by 10%. Produce the following plot:



j)From the data, it would appear that perhaps indeed Chlamydia incidence has been dramatically rising in the US over the past couple of decades. But take another look at the Discussion in the CDC study… is there another explanation for the patterns we see? Is our model valid if those explanations are true?