**AML 612 Spring 2017 Homework #3**

**Submit all files to** **smtowers@asu.edu****.**

**Due Wed March 1st, 2017 at noon.**

**Please submit with name format hwk3\_<first name>\_<initial of last name> Please provide your R file, and a Word file that gives the output to your R screen, plots, etc. The prospectus for your term project should be in a stand-alone latex document. Do not put your name in this latex document, and give it the name prospectus.tex (with compiled pdf prospectus.pdf)**

**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)**

By now you have read several papers regarding compartmental models related to topics that interest you.

Prepare a short prospectus describing an idea for a term project that involves some source of data (disease data, population data, etc… whatever you like) and a relatively simple compartmental model that can be fit to that data set. You can either extract the source of data using DataThief from one or more of the papers you’ve read, or look for sources of data online, or perhaps get ideas from the course web page describing lots of sources of free online sources of data <http://sherrytowers.com/2012/04/03/finding-sources-of-data-free-online-data/>

Compartmental models can be used to simulate many things. Disease is of course an example we’ve discussed many times in class. You can also simulate the spread of ideas as an infectious disease… these “ideas” can include things like the idea that committing a crime might be a good idea, or that smoking is desirable, or being obese is OK… having friends who are criminals, smokers, or obese can “infect” you with the idea to do the same.

The spread of memes in social media can also be simulated with a compartmental model. Or the number of hits over time on a YouTube video that goes viral. There are many, many things other than actual diseases that can be simulated with compartmental models for infectious disease.

Then there is population biology, where you can use compartmental models to simulate the change in populations, predator prey systems, wildlife management protocols, fisheries, etc.

In your prospectus (written in Latex, with bibtex references), give a few sentences **motivating** your proposed project, and then go on to describe your proposed **objective** (remember that “motive” is a description of why someone should be interested in your project, and why what has been done in the past is insufficient to really solve or understand the issue… “objective” is what you plan to do). Discuss whether this topic has been studied before.

Thoroughly describe your sources of data giving links and/or references, and a plot of the data.

Describe your proposed model, including a compartmental flow diagram and the model equations. Describe which parameters are known from the literature (give references!) and which must be obtained by fitting to the data.

Note that I’m not looking for a complicated model here… consider at most fitting for three model parameters.

Also note that you don’t have to do any of the model fitting at this point… you are just proposing a potential group project and describing your data and model.

After I’ve gone through them all once you’ve handed them in, the PDF files of these prospectus proposals will be circulated to all the other students in the class. Students will review the prospectuses, and also rank their top four choices of projects to work on. Based on student rankings (combined with my own opinions of which projects are likely to be feasible based on the data and model presented) I will assign you all to project groups. There is no guarantee that your project proposal will be chosen for a project group, but if it is, you will of course be in that group.

**Question 2)**

Norovirus is a common cause of gastroenteritis. The virus is transmitted by fecally contaminated food or water, by person-to-person contact, and via aerosolization of vomited virus and subsequent contamination of surfaces.

Norovirus infection is characterized by nausea, vomiting, watery diarrhea, abdominal pain, and in some cases, loss of taste. A person usually develops symptoms of gastroenteritis 12 to 48 hours after being exposed to norovirus. General lethargy, weakness, muscle aches, headaches, and low-grade fevers may occur. The disease is usually self-limiting, and severe illness is rare, except in the elderly and immunocompromised. Most who contract it make a full recovery within two to three days.

After infection, immunity to norovirus is usually incomplete and temporary, and all such immunity is gone after two years. Outbreaks of norovirus infection often occur in closed or semi-closed communities, such as long-term care facilities, overnight camps, hospitals, schools, prisons, clubs, dormitories, and cruise ships, where the infection spreads very rapidly either by person-to-person transmission, or through contaminated food. Many norovirus outbreaks have been traced to food that was handled by one infected person.

On a cruise ship, the staff normally work for weeks at a time, and live aboard the ship. Thus, while the passengers come and go every week (or however long the usual cruise lasts for that ship), the same staff remain.

Here we will study the number of identified cases of norovirus in passengers and crew over several cruises of a cruise ship that had 999 crew members <https://wwwnc.cdc.gov/eid/article/11/1/04-0434_article#r4>



In cruise number one, there were 2,318 passengers. In cruise #2, assume that there were 2,400 passengers. The 999 crew remain the same throughout all cruises. They believe that the outbreak started with one infected passenger on day 0 of the first cruise.

Here is the model I propose for these data:



The model has two classes of people (i=1,2), with the first being passengers, and the other being crew, with contact matrix C\_ij. Recall if the relative fractions in the population of the classes of people are f\_i, then the contact matrix has to satisfy reciprocity, which means that the total number of contacts that the crew makes with passengers has to equal the total number of contacts passengers make with crew. This means that f\_1\*C\_12 = f\_2\*C\_21.

We will assume for norovirus that the incubation period is 1.5 days, and the infectious period is also 1.5 days (see <https://www.cdc.gov/hai/pdfs/norovirus/229110-anorocasefactsheet508.pdf>)

1. What are the assumptions of this model?
2. Norovirus immunity upon recovery is not permanent and fully wanes within two years. Given this, do you think that our model adequately describes the dynamics of the spread of the virus on the cruise ship? Why?
3. Read the post <http://sherrytowers.com/2012/12/11/sir-model-with-age-classes/> Note that the R0 of an SEIR model is the same as that of an SIR model. We will assume in this analysis that beta\_1=beta\_2

Let’s assume that C\_11+C\_12 = 1 (and thus beta is identified as the number of contacts per day sufficient to transmit infection made by passengers). Due to reciprocity C\_21 = C\_12\*f\_1/f\_2. You will do a fit to the data, fitting for the following things:

* 1. The reproduction number R0
	2. The fraction of total contacts passengers make with other passengers, C\_11 (we can then calculate C\_12=1-C\_11, and C\_21=f\_1\*C\_12/f\_2)
	3. C\_22

Write the R code to solve the set of differential equations, and read in the data file <http://www.sherrytowers.com/cruise_norovirus_outbreak_data.csv>

Have the code use the Monte Carlo method to randomly sample different hypotheses for R0, C\_11 and C\_22 from uniform distributions in appropriate ranges, and do a Least Squares fit to the data from the first cruise for both the passengers and the crew (there is a column in the data file called “cruise number”).

The least squares statistic should be formed from the least squares statistic comparing the model predicted incidence in the passengers to the passenger data, added with the least squares statistic comparing the model predicted incidence in the crew to the crew data (ie; you are finding the combined best-fit to the data sets).

The initial conditions for the model at day 0 should be one passenger infected, and everyone else, including all the crew, susceptible. There are 2,318 passengers in the first cruise, and 999 crew.

Norovirus is known to have a significant fraction of cases that are asymptomatic, plus there is incentive for passengers and crew not to report being sick (captains of ships have the authority to confine passengers to their quarters) so there may be additional reasons people would avoid being counted as a case even if they are showing some symptoms. Thus, the number of cases observed is likely a fraction of the total actual number of cases. The sum of the model predictions for the incidence in the passengers and crew thus have to be normalized to the sum of observed incidence.

Do 10,000 Monte Carlo iterations, and re-create the following plot showing the best-fit for the passenger and crew data, and the results of the Monte Carlo iterations. Use a different colour scheme than that shown here. Note that because your random seed will be different than mine, your results might be slightly different, but the central best-fit values should be quite close to those shown here.



1. Now re-do the fit, but this time simultaneously for both the first and second cruises. Solve the system of equations for the first cruise just as above. For the second cruise, use initial conditions for the crew equal to the state of the crew at the end of the first cruise, and assume that all 2,400 of the new passengers are susceptible on the first day. The least squares statistic for the passenger data is thus the model predictions for the passenger incidence in the first and second cruise chained together, compared to the observed passenger incidence over those two weeks. And similarly for the least squares statistic for the crew data. Add these two least squares statistics to obtain the total least squares statistic for the joint fit, and do 10,000 iterations to produce the following plot with a different colour scheme:



1. what are the assumptions of the Least Squares fitting method? Which of those assumptions are met in these data?