**AML 610 Fall 2014 Homework #5**

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

**Due Wed Oct 8th, 2014 at noon.**

**Please submit with name format hwk5\_<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 many papers regarding compartmental models.

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. Then describe your sources of data (give links and/or references) and 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.

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 all then 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)**

The Boston Childrens’ Hospital has been maintaining a compilation of the number of case counts of Ebola in the ongoing West African outbreak: <http://healthmap.org/ebola/#timeline>

I’ve downloaded this data up to Sept 22, and compiled a summary of the number of case counts over time for the three primarily affected countries: Sierra Leone, Guinea, and Liberia. Note that because the reported data are cumulative case counts, they aren’t appropriate for fitting because the data violate the assumption of our fitting methods that the stochasticity in the data is independent from point to point… we thus need to calculate the number of new cases each report by taking the number of cumulative cases in the nth report, and subtract the number of cumulative cases in the (n-1)th report.

I went ahead and did this calculation for the number of new cases for each report, and put the data in the file [www.sherrytowers.com/west\_africa\_ebola\_data.csv](http://www.sherrytowers.com/west_africa_ebola_data.csv)

In the file, the day of the report is given relative to July 1st, 2014. (so day\_of\_report=2 means that report was on July 3rd). If you look at the first line of the file, you will so that there were, for instance, 20 new cases of Ebola recorded in West Africa between July 1st to July 3rd .

a) Write the R code to read in the file and plot the incidence vs the reporting date. You should get a plot that looks like this:



b) Read the recent paper written by myself and two co-authors: “Temporal Variations in the Effective Reproduction Number of the 2014 West Africa Ebola Outbreak”, at <http://currents.plos.org/outbreaks/article/temporal-variations-in-the-effective-reproduction-number-of-the-2014-west-africa-ebola-outbreak/>

Although I don’t specifically mention it in the paper, to obtain estimates of the rate of exponential rise in cases, I used the parameter sweep method we have been discussing in this course.

From the results presented in the paper, state what the estimated exponential rise was for the combined West Africa data going up to Sep 8th.

c) In this homework, you aren’t going to fit a compartmental model to this data (although we could certainly go ahead and fit the parameters of an SEIR model to this data, and fit for R0). Rather, to keep things simple, we are going to assume that we are still in the initial exponential growth phase of the outbreak. Note that if you *know* you are probably still in the exponential rise of epidemic growth, it is always easier and faster to fit an exponential curve, because you have an analytic expression for your model, rather than having to deal with all the computational overhead of having to numerically solve an SEIR model over and over.

You will use Equation 1 in that paper (along with the normalization condition right below it) as your model: the exponential rate of rise, rho, is the parameter we will fit for. Specifically, the model is

Ypred = exp(rho\*tmax)-exp(rho\*tmin)

and then you normalize Ypred to sum to the data by setting

Ypred = Ypred\*sum(Ydata)/sum(Ypred)

The quantity “tmax” is the day of the report for each data point. For “tmin” use the day of the previous report (for the first data point assume tmin=0)

Write the R code to do 10,000 iterations where you randomly uniformly sample rho in some appropriate range (you may have to try a few different ranges to find one that isn’t too wide, and also contains the best fit value).

For each hypothesized value of rho, calculated Ypred. From this value of Ypred and the data Ydata, calculate the Pearson chi^2 statistic.

Plot the values of the Pearson chi^2 statistic vs the rho hypotheses, and also plot your best fit model overlaid on the data. You should get a plot that looks something like this:



d) The data (and our best-fit model prediction) appear to jump up and down a lot because there is a lot of variation in time between Ebola situation updates. More new cases will accrue if there is 7 days since the last update, compared to if there has only been 2 days since the last update. To correct for this, write the R code to plot both your data and your best fit model prediction, divided by the time since the last update… this is now plotting the average number of new cases per day during that time period. Your plot should look something like this:



Comment on whether the data seem over or under-dispersed compared to Poisson stochasticity. Do you think Poisson likelihood was a good choice for this data? What might other possible appropriate choices be?

e) from the results of c) estimate the best-fit value of rho, and it’s one standard deviation uncertainty.

* Is your best fit value of rho bigger or smaller than the one in the paper? Does the change in value (in light of the more recent data you are using in this homework, compared to the more out-of-date data in the paper) bode well or ill for the near-term progression of the outbreak?
* Calculate the 1std dev width of your confidence interval (note: to compare to the 1std dev width of the confidence intervals quoted in the paper, take your maximum confidence limit, minus your minimum, and divide by 2).
* How many times smaller is this than the 1std dev width of the confidence interval quoted in the paper (ie; the number after the “±” sign)?
* Why do you think your confidence interval is so much narrower?
* Which confidence intervals do you think are probably the most correct; the ones derived with the Poisson negative log-likelihood, or the likelihood used in the paper?
* Why did the paper use the likelihood that it did?

f) Using eqn 2 in that paper, calculate the estimate of R0 for an SEIR model. Assume that 1/kappa=10 days, and 1/gamma=10 days. Is the value of R0 you get more or less in agreement with the values obtained by other analyses of this data?