**AML 612 Spring 2017 Homework #5**

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

**The written parts are due Wed March 31st, 2017 at noon.**

**The group oral presentations will occur on Monday March 29th, and Wednesday, March 31st.**

**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.**

**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)**

In Homework #3, you fit a two-class SEIR model to norovirus outbreak data from a cruise ship. In that analysis, you used Least Squares fitting methods, which are inappropriate for that data because it is count data with widely varying counts per bin, and also many bins with low counts per bin.

In this analysis, you will switch to using likelihood methods by doing a Poisson likelihood fit and a Negative Binomial likelihood fit, and also assess one standard deviation confidence intervals on the reproduction number of the outbreak.

In the original analysis, you used –delta(S) to assess the number of newly identified cases per day. While this would be the correct way to do it for an SIR model in the absence of births and deaths, in reality, this is not the best estimate for an SEIR model, because there is an exposed period during which time people have no symptoms, and it is only when they move to the infectious (and symptomatic) compartment that they are counted (in the case of norovirus, the average time spent in the E class is short, so the –delta(S) method will yield a reasonably ball-park estimate of the number of newly identified cases per day).

In order to properly assess the number of newly identified cases per day when using an SEIR model, create another model “compartment” to be solved in your deSolve methods that counts the total number of people who have flowed into the I compartment (call it newI, for instance). Thus, the differential equation for newI would look like

 dnewI\_dt = +kappa\*E

Once you have the numerical solutions for the S, E, I, R, and newI compartments, the number of newly identified cases per day will be +delta(newI). You will need to do this process for both the crew and the passengers for this analysis.

Normalized the predicted incidence of the passengers to the observed passenger incidence. And similarly for the crew. Assume the time of introduction of the virus was the day before cases were first identified.

In the following, you will be using the fmin+1/2 method, and the weighted mean method, to assess the estimates of the best-fit parameters, and their one standard deviation uncertainty (and, equivalently, the one standard deviation confidence interval). In the following you will be doing Poisson and Negative Binomial likelihood fits to estimate R0, C\_11, and C\_22.

You will produce plots of the negative log likelihood vs the sampled parameter hypotheses, with the best-fit value and one standard deviation uncertainty for each parameter shown in the title of plot of the likelihood vs hypotheses for that particular parameter. Also show a set of arrows depicting the one standard deviation confidence interval when the likelihood is equal to fmin+1/2. When you make these plot, use a different colour scheme than I use below, and also do enough Monte Carlo iterations to ensure that the plots are well-populated with points (for instance, my plots for C\_11 and R0 could stand to be better populated with points for the fmin+1/2 method…. 10,000 MC iterations wasn’t enough in this case). Also, note that your central value and one standard deviation CI will be somewhat different than mine because your random seed will be different. Notice that in the plot titles I put the best-fit estimates from both the fmin+1/2 method, and from the weighted mean method, and the estimate of the one standard deviation uncertainty on those estimates, and their estimated one standard deviation CI. Your plots should similarly have these values in the titles over the plots, and the values you get should be similar to mine. Note that the one standard deviation CI for the weighted mean method goes from the best fit value from that method minus the one std dev uncertainty estimate, to the best fit plus the one std dev uncertainty (ie; it is symmetric about the best fit). The one std dev uncertainty from the fmin+1/2 method is not necessarily symmetric about the best fit.

Remember: for the weighted mean method to work, you have to **uniformly** sample parameter hypotheses (no preferential sampling near the best-fit using rnorm), and the range you sample has to cover at least 3 or 4 standard deviations. For the fmin+1/2 method to work, you have to have the plots quite populated with points near the best fit values.

This exercise should make something clear: with the weighted mean method, you need to randomly sample uniformly over a broad area, so you lose computational efficiency because you can’t preferentially sample using rnorm close to the best-fit value. With the fmin+1/2 method you can use preferential sampling, but you really have to make sure your plots are well populated with points, so the procedure is computationally intensive.

1. For the Poisson likelihood fit, produce the following plot, with a different colour scheme than I used here, and also do enough Monte Carlo iterations to ensure that the plots are well-populated with points (for instance, my plots should be better populated with points for the fmin+1/2 method…. 10,000 MC iterations wasn’t enough in this case). Also, note that your central value and one standard deviation CI’s will be somewhat different than mine because your random seed will be different. The titles on each plot must contain your best fit estimates, and the estimates of the one standard deviation uncertainty and confidence interval for both the fmin+1/2 and weighted mean methods.



1. Do a Negative Binomial likelihood fit, and produce a plot similar to this one, with a different colour scheme. Recall that alpha is the over-dispersion parameter used in the Negative Binomial likelihood. Your plots should be better populated with points than these ones. Again, your one standard deviation confidence intervals and best-fit values will be somewhat different than the ones I obtain.

The best-fit values of R0 and C\_11 are quite different than those obtained in the Poisson likelihood fit. Why do you think that is?



1. Which appears to be the most appropriate estimation method for these data; the fmin+1/2 method, or the weighted mean method? Which is the most appropriate fit statistic: the Poisson or the Negative Binomial log likelihoods?
2. For the Introduction section of a paper describing this analysis, write two or three paragraphs that motivate the study of norovirus; starting general, and moving on to cruise ships specifically. Cite at least three papers. First describe the epidemiology of norovirus, the number of hospitalisations each year, cost to the economy, etc, etc. Then move on specifically to discussing norovirus in cruise ships (number people affected, costs, etc).

Then write a paragraph or two with at least two references that describe what has been done in the past to model norovirus outbreaks. Then write a few sentences describing the objective of this analysis and what is novel about it (to the best of my knowledge, our model with crew and passengers is novel, but I would like each of you to do literature searches related to norovirus models to see if this is indeed the case… **25 bonus points over the possible 100 points you can get in this homework if you can find a past modelling analysis that includes crew and passengers in the model**). Note: simply fitting a model to norovirus outbreak data provides useful info about the R0, but what is novel about our analysis is not that we estimate the R0… the use of our model with passengers and crew can help to inform control strategies (like, for instance, enhanced control measures aimed at reducing crew-to-crew transmission, or crew-to-passenger transmission, for example). Just doing a fitting analysis of a model to data is not an objective in and of itself… the point is then to use the fitted model to better understand control strategies (if modelling a disease) or harvesting/control/optimization strategies for biological populations, etc, etc.

**Question 2)**

Once I assign the project groups, each project group will prepare a short presentation (15 minutes, plus 5 to 10 minutes for questions) on an analysis that was cited as a background reference in the original project prospectus. *You will be giving this presentation with the same detail as if you were the people who had actually done the analysis*. Thus, the presentation is expected to have an introduction to the topic describing the motivation and prior previous work, moving on to the objective, detailed methods, detailed results, and discussion of how the work contributes to the body of literature on the topic. The project group will tag-team the presentation.

All students in the audience will be expected to pose one question each to the presentors during the question period after each of the talks.

The project groups will give their presentations on March 29th and March 31st during class time. I’ll tell each group which date they’ll be presenting once I assign the project groups.