Name _____

Problem Set 3
HEB 1333
Fall 2012

An important note before starting this problem set:
If you have a version of R older than 2.15 (your version number will appear at the top of the console text on startup), you will need to update R for this problem set.

Welcome back to the HEB 1333 lab. Your conclusions about the epidemiology of HEBV from our last meeting have proved to be crucial in controlling this virus within the Boston population! We have averted what could have turned out to be a public health nightmare, but there's still much more to be done- so suit up and get ready for part two of our analysis of this virus…

This time, we're going to be looking at the spread of HEBV over transmission networks and attempting to determine risk factors for contracting HEB fever. As before, all of the data relevant to this analysis can be found on the course website, this time in six data files ("epidemiology_data.csv", "sociomatrix.csv", "undirHEBV.csv", "obs_spread.csv", "obs_spread_symptomatic_ONLY.csv", and "survey_data.csv"). All analyses will again be conducted in R, and this time, you'll have a little more freedom to do some data exploration. Make sure to also download all of the R script files that go along with this problem set ("simulation.r", "NBDA_basic.r", "NBDA_extended.r", and "survey_starter.r").

Remember before starting any of your work for this problem set to change your working directory (hint: `setwd("<folderpath>")`) to the folder on your computer where you have saved all of the data files and R script files.

1.a) Begin by importing to R "epidemiology_data.csv"– a file that we worked extensively with in the last problem set– and name the dataset whatever you'd like. Let's start with a warm-up question from this dataset: how many individuals are there in the HEB 1333 population?

_____

1.b) Write the R code that you used to obtain this number:

1.c) We discovered in our first analysis that the R_0 of HEBV was approximately 2.97, but another important variable to know for simulating and understanding the spread of a disease is infectious period. Using the data that you've just imported, calculate the average infectious period for an individual infected with HEBV (hint: determine how long each individual in the population was infectious [from infection start to infection end] and then average all of these numbers using `mean()`); also write the R code that you used to obtain this average:

Name _____

Because of the excellent data collection by our team of researchers, we have obtained unrealistically accurate tracking of the epidemic and *all* infected individuals. In addition to this transmission data, we've also obtained a social network of contact among individuals in the study population. We will use this information to structure simulated spreads of the parasite through the population. In the next few sections, we'll see how our "unrealistically complete" data compares with more realistic estimates of infection obtained through self-reporting of symptoms, and how real-world constraints affect our understanding of the epidemic.

2.a) We're extremely curious to know how HEBV is transmitted– either socially, environmentally, or otherwise– because such knowledge can have important implications for disease control and isolation of particularly high-risk groups. We've developed a simulation for different transmission modes of the parasite, and we'd like you to test out our models. Run the R script file for the simulation by typing `source("simulation.r")` into the R console. This model will simulate the spread of HEBV through our previously compiled social network. Follow the prompts, inputting the R_0 and the average infectious period estimate that you calculated in question 1, and then choose the option for "non-social transmission" of the parasite. In watching the progression of the disease, do you think that this simulation accurately models the actual spread of the parasite? (Network animations and SIR graphs of the actual spread are available in the HEBV Case File on the course website for comparison.)

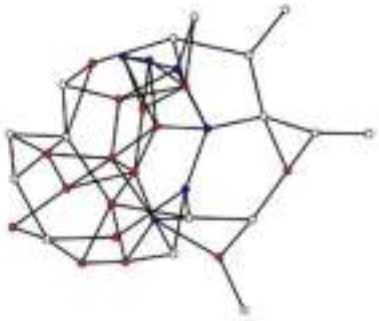
Once the animation stops, make sure to name the output file of this simulation something that you will remember (probably something with "non-social" in the name): we'll be needing it again shortly.

2.b) Repeat the simulation exactly as in 2.a, but this time, run the "social transmission" simulation. What graphical differences do you notice between this run of the model and the non-social model? Also compare the graphs generated by this model to the ones in the HEBV Case File.

As before, name the output file of this simulation with a catchy title (this time, probably include "social" in the name).

Name _____

Network-based Diffusion Analysis (NBDA)



Mathias Franz, a researcher at Duke University (who was also a former Ph.D. student of Professor Nunn) has come up with a method to distinguish social transmission from non-social transmission of a trait within a socially-structured population (or, social network). The method uses what is called Akaike's Information Criterion (or AIC, for short) to determine whether one model fits the observed data better than another. A rule of thumb when evaluating AIC values is that the *lowest* value is usually indicative of the better fit, and this is even truer when one AIC value is at least 2 points lower than any other value (just a convention). Luckily, Franz's method also provides Akaike weights, which are even simpler to interpret: each Akaike weight for a given analysis can be interpreted as the probability that a given model is the best fitting model. In this case, *higher* Akaike weights are more indicative of a better model fit.

3.a) We'll start by analyzing the output from our earlier simulations. This is useful because we know how the data were generated, and we can therefore see how well the NBDA works: it should favor non-social transmission models for the non-socially generated data, and social transmission models for the socially generated data in our simulations. Run the R script file for an NBDA comparison between pure social and pure non-social model fits by typing `source("NBDA_basic.r")` into the R console. When prompted for a sociomatrix file (this is how social network data are often stored, as pairwise interactions), type "sociomatrix.csv", and when prompted for a spread data file, first input your personally named "non-social" file. Which of the two models were best supported by NBDA? What was the Akaike weight for this fit? How confident are you in the fit of this model?

3.b) Repeat the directions for question 3.a, but this time for your "social" spread output. Which of the two models were best supported by NBDA? What was the Akaike weight for this fit? How confident are you in the fit of this model?

Name _____

3.c) Based on your observations of the spread graphs in the HEBV Case File in comparison to the two simulated spreads, which of the models do you predict will be most strongly supported by NBDA for the actual spread of HEBV in the population? Why?

3.d) Now for the moment of truth: run the NBDA code again by typing `source("NBDA_basic.r")` into the R console. Again, choose "sociomatrix.csv" for the sociomatrix option, but this time, choose "obs_spread.csv" for the spread data file; this spread data file includes the time at which each member of the class was infected with HEBV. Which of the two models were best supported by NBDA? What was the Akaike weight for this fit?

3.e) Was your prediction correct? If your prediction was incorrect, why do you think this was so? (Note: it is *absolutely acceptable* to predict something incorrectly in science: it happens all the time, and this is often where the most interesting and creative new hypotheses come from.)

4.a) The basic NBDA approach that we were using was slightly narrow in it's approach: it only allowed your model to be purely social or purely non-social, which can be a problem. In an extension of the NBDA approach, we can test for a combination of social and non-social transmission versus pure non-social transmission. Run this model by typing `source("NBDA_extended.r")` into the R console. Again, choose "sociomatrix.csv" and "obs_spread.csv" for the input files. What do you find this time? If you found a result overturning your conclusions from question 3.d, why do you think this was? Consider carefully how HEBV spread through the population; was it purely socially contagious, or were there some individuals who appear to have just infected other individuals at random?

Name _____

4.b) Do you think that your findings in question 4.a have broader implications for outbreaks of diseases across the world? Do all infected individuals always know each other?

4.c) So far, we've been working with our idealized dataset of perfect coverage of all infected individuals. What happens when we restrict our analyses to only those individuals who self reported illness to course "medical staff"? To start with the basic model, type `source("NBDA_basic.r")` into the R console. Again use "sociomatrix.csv", but this time, use "obs_spread_symptomatic_ONLY.csv" as the spread data file (a spread file which only includes infection data for symptomatic individuals). How do these results compare to those of question 3.d? Are you surprised that these results were similar? Why or why not?

4.d) Again, run the extended NBDA using `source("NBDA_extended.r")`, but this time using the dataset on infection times of symptomatic individuals only ("obs_spread_symptomatic_ONLY.csv"). Do you notice any difference in the support for models between the extended NBDA approach here and that in your answer to question 4.a? What do these results suggest for self-reporting of illness/infection or public health monitoring of a disease outbreak?

4.e) As a public health consultant to HEB 1333 Labs, Inc., how would your suggestions differ (if at all) for control of HEBV given the results of both 4.a and 4.c? Why would you make these suggestions, and are you confident that they would work in controlling the spread of the virus?

Name _____

5.a) How do certain risk factors affect the time of infection by HEBV? If there are certain risk factors that we can show are significant in increasing an individual's chance of becoming infected, then we may be able to target our vaccinations and become even more efficient in controlling the virus' spread. We've attached a short R script file to get you started with this data exploration, just run the code `source("survey_starter.r")` in your R console and you will have a fully prepped R dataframe named "`survey`". Explore this dataframe using the R commands `head()` and `summary()`. What variables are included in this dataset? What do each of these commands tell you about these variables (in general)?

5.b) Let's explore some variation in the data using histograms. We can make histograms in R using the command `hist(<variable>)`. Produce a histogram of the number of social connections of each individual in the class (this includes both the ties that each individual reported to others and the number of ties that others reported to them in the "HEBV Aftermath Survey"). Print this graph out and include it with your problem set submission. What kind of information can you glean from this graph- does there appear to be substantial variation in the number of social connections per individual?

5.c) Let's also look at some relationships between different variables. We can do this graphically by producing a wide variety of different graphs with R's `plot()` function. Plotting in R can use numeric or non-numeric inputs and follows the basic syntax of: `plot(<response variable> ~ <predictor variable>)`. Start off by plotting infection time as the response variable and number of social connections as the predictor variable. Print out and attach the graph. What relationship do you see between these variables?

5.d) Plotting can also produce boxplots, which allow you to compare the distributions of data in different groups. Make a boxplot with infection time again as the response variable and sex as the predictor variable. Print out and attach the graph. What relationship do you see between these variables?

Name _____

6.a) As we noticed before, when we restrict our data to only those who reported symptoms, we can get very different outcomes from the full dataset. We've prepared another R dataframe for you with only the symptomatic individuals, and this dataframe is titled "`symptomatic.Y`". How does the histogram of connections in the symptomatic subset compare to the one produced in question 5.b? Is it much different? Print off this histogram and also include it with your submission.

6.b) Let's also compare our new graph of infection time versus social connections for the symptomatic subset against our graph for the full dataset. How does this graph for the symptomatic subset compare to the one produced in question 5.c? Print off the graph and include it with all the others.

6.c) Finally, let's compare a boxplot of infection time by sex for the symptomatic subset against our boxplot for the full dataset. How does this graph for the symptomatic subset compare to the one produced in question 5.d? Print off the graph and include it with all the others.

6.d) Based on your results from questions 5 and 6, do you think that symptomatic reporting had a substantial effect on conclusions about risk factors in contracting HEBV? Do you think that symptomatic reporting had as large of an effect on risk factor analyses as it did on your network transmission analyses (NBDA) in questions 3 and 4? Why do you think this was?