MTBI 2013: Dr Towers’ Homework #2

Please hand in a Word document with your R code, and copies of the plots output by your R code. All your R code must be well commented throughout to explain what it is doing, and also must contain a proper header comment, as described in <http://sherrytowers.com/2012/12/14/good-programming-practices-in-any-language/>

Poorly commented code will be returned for a redo.

All plots must exhibit the qualities of a good plot, as described in <http://sherrytowers.com/2013/01/04/good-practices-in-producing-plots/>

1. In homework#1, question 3, you were asked to use DataThief to extract data from a figure in a paper. Read the file into R and plot the data with appropriately labeled axes. When submitting the homework, on one page of the Word document give a screenshot of the figure from the paper, followed by your R figure copied and pasted into the Word file.
2. Go to the CDC Wonder mortality website and download mortality data by year for a state of your choice and for your choice of ICD-10 cause-of-death code. Request age-adjusted rates. Edit the data file such that it can be read into R (remember; get rid of the “Notes” column, remove spaces from column names, and remove the annotations at the bottom of the file). Write the R code to read the file into R and plot the number of deaths per year with appropriately labeled axes. Each student should submit a unique set of data. Copy the figure and paste into the Word document. Copy and paste your R code in to the Word document.
3. From the CDC Fluview weekly archives for the 2012-13 influenza season (<http://www.cdc.gov/flu/weekly/>) you can download the influenza confirmed weekly case count data by clicking on the “View Chart Data” link below the first figure. I’ve taken this data and put it into the file [www.sherrytowers.com/flu\_2012\_2013.txt](http://www.sherrytowers.com/flu_2012_2013.txt) and modified it such that the weeks are expressed relative to Jan 1st 2012. Read this file into R and subset the data to only include data before week 60. Plot the number of confirmed influenza B cases by week from the sub-setted data set, with appropriately labeled axes, and as a line plot (not points). Now overlay the plot of AH3 cases by week, also as a line plot, and with a different color. Overlay the AH1\_2009 cases by week, as a line plot with yet another color. Adjust the line widths to make a visually pleasing plot. Give the plot a main title that is descriptive. Note that you may have to adjust the range of the y axis to fit all curves on the plot. Put a legend on the plot indicating what each line color represents. Copy the figure and paste it into the Word document. Copy and paste your R code that produced the paper into the Word document. Your plot should look like this:



1. In www.sherrytowers.com/ MTBI\_hwk2\_question4\_data.txt you will find simulated prevalence data (number of infected people over time) from a hypothetical epidemic. People who catch this hypothetical disease rarely die from it, and once they recover they are permanently immune (ie; an SIR model is appropriate for this disease), Write the R code to read this data in, and plot prevalence vs day, with appropriate labels for the axes and appropriate line width. From observational studies, epidemiologists know that the recovery rate of this disease is gamma=1/4 days^{-1}, but they do not know the reproduction number R0, which for an SIR model is the transmission rate, beta, divided by the recovery rate. The epidemiologists also don’t know the initial number of infected people that entered the population at day 1, I\_0, but they believe that the population (except for those initial number of infectives) was completely susceptible to this new disease. They also know that the population size is N=100,000Write the R code to do 5000 iterations, where in each iteration you randomly sample hypotheses for the values of R0 and the initial number of infectives I\_0, Randomly sample both parameters from a uniform distribution; for the range of the R0, try 2.98 to 3.02. For the range of the I\_0 try 95 to 105. For each iteration, calculate the predicted number of infected people each day, I, using an SIR model calculated using the (R0, I\_0) hypothesis. Compare this model to the simulated data, and calculate the least squares statistic for each R0,I\_0 hypothesis. Fill vector with the R0, I\_0, and least squared statistics from the iterations.
2. Plot the least squared statistic vs R0. Only plot values for which the least squares statistic is less than 7.5e5. On the same page, plot the least squared statistic vs I\_0, again only plotting points for which the least squares statistic is less than 1e6. What is the estimated best-fit value of R0? What is the best-fit value of I\_0?
3. Repeat questions 5 to 6, except calculate the Pearson chi-squared statistic. Only plot values for which the Pearson chi-squared statistic is less than 70. What is the best-fit value of R0? What is the best-fit value of I0\_? Bonus points: what is your estimated one-standard deviation uncertainty on R0 and I\_0?

The plots produced by questions 5 through 7 should look a lot like the following plot (although yours will look slightly different than this because your random seed will have been different than mine):

