Jared Yu

HW 2

STA 243

1) The first problem dealt with the method of Simulated Annealing. The method of Simulated Annealing is such that it included a level of probability in the process of optimization so that the randomness can help the algorithm to better discover the global maximum of a set of data.

To replicate this process in R, it required the use of different mathematical functions that were to be transformed into code. First, the distance matrix was placed into a large vector where it could be accessed readily. Next, the , and distance function are created and put into R code. Then, the functions are assembled inside a larger function that acts as the algorithm for the program. The goal of the program is to track the previous candidates and each updating candidate to check if an improvement in the distance between each travel time is found.

Since an ideal distance of 17 was described as the answer, different attempts were made to locate this solution. Initially, different parameters were attempted, such as various values for and . The strategy was to pinpoint an ideal range within the parameters after several tests to try and discover which set of parameters would assist in finding the solution 17. However, after only coming close at most to 21, and not being able to determine a suitable range which seemed accurate this strategy was left behind. Speaking with the professor, it was learned that the ideal parameters were already provided. Therefore, the problem seemed to be within the code. After creating an improved version of the algorithm, it was possible to use a while loop rather than a for loops and sapply() functions to obtain the solution.

After having re-done some of the code so that it could be run in a more efficient manner, the function could be executed in a simpler and cleaner style. This may have been the reason for the difference of the original solution of 21 rather than 17. Additionally, a new seed was set to obtain this answer.

2) The second problem utilizes Genetic Algorithms to solve a problem which can't be approach from the traditional Linear Regression model. Genetic Algorithms work by including a method inspired by evolution taught in Biology classrooms.

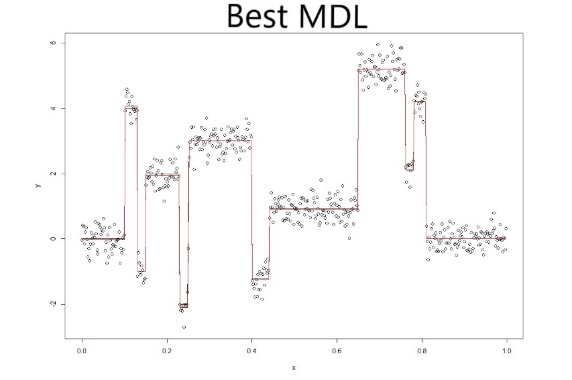
To fit the points that are given in the question, a piecewise function must first be estimated. The piecewise function is the construction of various separate functions which are themselves averages of the data within each interval.

Two processes are used to determine the offspring which will consist of future generations. One of the processes is Crossover, where both parents are mixed to generate a new offspring of the same length as one of the parents. The other process is Mutation, where one of the parents are randomly selected, and some of its chromosomes are altered.

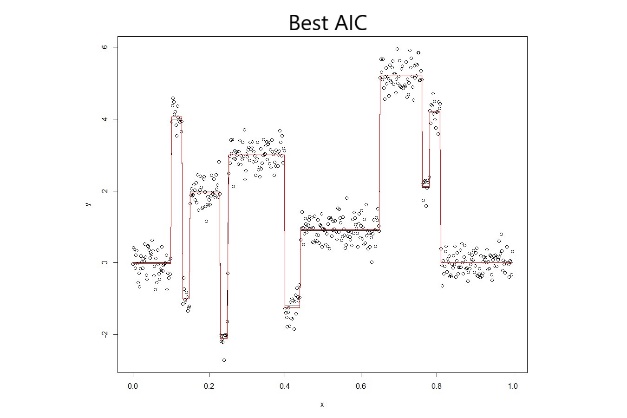
Additionally, the method for testing the fitness of each population can be either Minimum Description Length Principle (MDL) or Akaike Information Criterion (AIC). One method is used at a time to determine the best fit chromosome of each population, and the best chromosomes are measured against future populations to determine if any improvement has been made.

For the R code, several smaller functions were first created. In the end they were assembled together in a large function. It can test a certain dataset along with a method for checking the fitness of each successive generation. The methods possible to use are MDL and AIC that were mentioned earlier.

From the program, the MDL output gave a graph which very closely mimicked the data points on the graph. The lines are much straighter than when tested with smaller parameters for number of chromosomes. The resulting chromosome which gave this output is mainly consisting of 0’s and only a few 1’s in between.



The AIC output also mimicked the data points closely on the graph. The chromosomes which gave this output were like MDL in that they are mostly 0’s with some 1’s in between.



I think that this assignment showed me that there requires great patience in transferring mathematical formulas into R code. This task was best performed step by step, and then in the end combined into a general algorithm/function which could be run at once. The process allows for improvements when necessary to be made, and changes to happen that can lead to important solutions. Often simple tasks can appear rather difficult at first, but in hindsight were likely daunting due to never having attempted a similar task before. Examples include transferring the MDL and AIC functions to R. The greatest lesson learned was in the utilization of time management. Never in the past was there a necessity to allow a program to run over night to obtain a solution. This was the first time, and it really showed how actual work can take much longer than developing a possible solution on paper. It requires that when computer computation is used that we accommodate for the time necessary to derive certain conclusions. Therefore, computational statistics can be understood as a combination of a human developing statistical formulas, and a computer being utilized to obtain certain solutions.

Note: Two Excel spreadsheets are included which are the chromosomes for both the AIC and MDL solution. I’m not sure that they’re required, but I thought that it would help to show that my solutions worked out.