System Simulation, Fall 2018

Modeling Challenge Problem #3: Regression Modeling and Simulation of Old Faithful

In this problem set you will complete the following:

a) Build a regression model that can be used to predict the eruption times (when the geyser will erupt) of the Old Faithful Geyser.

b) Build a random number generator that can be used to generate estimates of how long the geyser will spend erupting (ejecting water and steam)

c) Use the products of #1 and #2 to build an event-driven simulation that simulates eruption times and durations over the course of a week

d) Compare your simulation derived data with actual data to assess the quality of your model

In short, you will be creating a complete simulation using a combination of tools you already know (event-driven simulation and custom random number generators) with a newly introduced tool (regression modeling).

Conceptually, your simulation MIGHT look like this:

```
world_clock = 0.0;
duration_of_eruption = eruption_duration_rng();
put eruption event in event_queue
while (world_clock < one week)
   { eruption_record = dequeue(event_queue);
     print time and duration of eruption to screen or file;
     world_clock += eruption_record.duration;
     time_to_next_eruption = world_clock +
                             eruption_time_predictive_model(eruption_record_duration);

     duration_of_next_eruption = eruption_duration_rng();
     put eruption event in event_queue;
   }
```

Note, there is more than one way to set up this simulation and I may have been a little fast and loose with my descriptions of how to compute times. Figuring out the actual timing calculations is an item for your group.

So, in small groups, please work on the following items in this order:

1) Using the data set old_faithful and any tools with which you feel comfortable, create a simple linear regression model of "time to next eruption". Your independent variable is the length of the current eruption in minutes. Your dependent variable is the length of

time, also in minutes, until the next eruption will occur. In your group, comment on the quality of the model using the R^2 and/or MSE (Mean Squared Error) criteria. Do you think this a good predictive model? Also, comment on whether your model is deterministic or stochastic. Could this be a problem? If so, under what conditions? Might there be a way to improve the model under those conditions? If so, how? When you are ready as a group to discuss these questions, call over the instructor.

2) Create a random number generator that outputs eruption durations (in minutes). You may use either rejection sampling or inverse transform sampling at your convenience and whim. When you have completed this task, use the RNG to create a population of eruption durations. Create a histogram of your generated data and compare to the relative frequency data of your original dataset of real eruptions. Are you happy with your RNG? Why or why not? When you have a group answer to this question, call the instructor for discussion.

3) Using the products of #1 and #2 above, create an event-driven simulation that will generate a simulated week's worth of eruption times and durations. Compare your simulation's generated data with the real data you started with. Do you believe that your simulation captures the essence of the real phenomenon? Why or why not? When you have an answer, call the instructor for discussion

4) You undoubtedly made a few modeling assumptions along the way. Create a list of the assumptions you made and discuss among yourselves what happens if each of these assumptions is violated. For each violation, does your confidence in the model go down? If so, how and why? When your group has completed this exercise, call your instructor for discussion.