Stat E139 Project Proposal: Predicting Marathon Finish Time Generally and by Subgroup
November 30, 2015

Project Team:

- Nathaniel Burbank nathaniel.burbank@gmail.com
- Pooja Singh goodwillyoga@gmail.com
- David Wihl davidwihl@gmail.com

Outline

We plan to analyze Boston and Chicago Marathon finish-time data from the years 2010 to 2015. First, we plan to build a multi-regression model that attempts to predict a runner's finish time based on their 5k split time (the number of minutes it took the runner to run the first 5 kilometers of the race), age and gender. Then, using various statistical tools we'll subset the data across criteria such as gender or age (and possibly other characteristics) to segment the population of runners into meaningful subpopulations that participate in the race. We'll first use ANOVA and other statistical techniques to identify the most meaningful subgroups. Then, with the groups in hand, we'll return to our original prediction question and analyse how much this added segmentation helps (or fails to help) us accurately predict a runner's marathon finish time.

We are also considering several other questions, that we may address depending on time:

- Attempt to predict runners who drop out of the race
- Analyze the effects of heat and running course by comparing runner performance between different years of the Boston Marathon and by contrasting the Chicago Marathon. Chicago features a much flatter course and does not require qualification.

Completed So Far

- Set up team organization (Trello board and github repository)
- Secured complete split time data for the three years (2010,2011,2013) that was made available by the researchers of a previous study analysing results from the Boston Marathon, as well as a separate dataset of finish time data for five years (2010-2015) directly from the BAA
- Downloaded historical weather data for Boston Marathon dates
- Decided on a toolset (R Markdown)
- Performed several ANOVA and Tukey comparisons along different groupings
- Performed initial regressions with polynomial transforms and measured residuals.

Remaining To Do

- Using ANOVA and other means such as kNN so as to cluster data into subgroups.
- Obtain finish time and split time data for the Chicago Marathon
- Improve regressions by both transforms and attempting different regression models such as Random Forest
- Perform regressions on subgroups and compare accuracy and residuals against general regression model.
- Write up conclusions

Challenges Faced

- The split time data we have access to has already been filtered to remove runners who started but did not complete the 2010 and 2011 races. This will make attempting to predict runners who won't finish difficult.
- Finish times are not linear with regards to split times, so some form of transformation will necessary