Case-Study 2: Modeling Runners' Times in the Cherry Blossom Race

(By Daniel Kaplan, Macalester College and Deborah Nolan, University of California, Berkeley)

2.1 Introduction

In this era of 'free and ubiquitous data,' there is tremendous potential in seeking out data to bring insight to a problem we are working on professionally or to a topic of personal interest. For example, we are interested in understanding how people's physical performance changes as they age. One source of data about this comes from road races. Hundreds of thousands of people participate in road races each year; the race organizers collect information about the runners' times and often publish individual-level data on the Web. These freely accessible data may provide us with insights to our question about performance and age.

One example of the many annual road races is the Cherry Blossom Ten Mile Run held in Washington D.C. in early April when the cherry trees are typically in bloom. The Cherry Blossom started in 1973 as a training run for elite runners who were planning to compete in the Boston Marathon. It has since grown in popularity and in 2012 nearly 17,000 runners ranging in age from 9 to 89 participated. The race has become so popular that entrants are chosen via a lottery or they guarantee a spot by raising $500 for an official race charity. After each year's race, the organizers publish the results at http://www.cherryblossom.org/. These data offer a tremendous resource for learning about the relationship between age and performance.

Overall guidance

You are free to produce any interesting analytics and visualizations that you want – ensure they are insightful and useful.

Furthermore, see if you do some modeling work such as performance as a linear function of age (you can assume any model but at this stage linear models are most preferable).