Guide to the Boston Marathon Webpage

Richard Smith

September 16, 2013

The file TIM.txt contains the full dataset for all 69,923 runners for 2013, 2011, 2010 combined. It's a plain text file and can be read into R with the command

TIM=read.table('TIM.txt',header=T)

The variables included in this file are:

- BibNum: Bib number
- Year: which year the participant ran the race (2013, 2011 or 2010)
- Age: Age on race day
- Gender1F2M: =1 if the participant was female, =2 if male
- StartHr, StartMin: 2013 participants only: the exact time the participant crossed the start line (StartHr is the hour, StatMin is minutes including decimal parts of a minute). This information was used in some analyses to determine whether it was possible for someone to have crossed the finish line before the bombs went off at 2:49 pm.
- K0-5: Split time from the start to 5km (minutes and decimal parts of a minute)
- K5-10: Split time from 5km to 10km
- K10-15: Split time from 10km to 15km
- K15-20: Split time from 15km to 20km
- K20-25: Split time from 20km to 25km
- K25-30: Split time from 25km to 30km

- K30-35: Split time from 30km to 35km
- K35-40: Split time from 35km to 40km
- K40-Fin: Split time from 40km to finish
- HalfMar: Split time at half marathon
- Age2014: For 2013 entrants, age on 4/21/2014 (date of next Boston marathon)

Throughout the file, NA is the missing value code, consistent with R.

Similarly, the file TIM1.txt contains 21,930 runners (derived from TIM.txt) who were either DNF in 2013 or finished in over 4 hours in 2013, 2011 or 2010. This is the file that was actually used in producing the projected finish times for 2013.

TIM2.txt is the validation dataset with 16,302 runners. The DNFs from 2013 were not included, but we artifically created DNFs among the other runners, so the the proportion of runners who dropped out at each stage of the validation dataset match those in the full dataset.

For those who prefer Excel files to plain text files, TIM.xlsx, TIM1.xlsx and TIM2.xlsx are the same three files reformatted using Excel.

TIM2.mat is the TIM2 file in Matlab format, which is used in the KNN and linear regression code files.

The file ProjectedTimes.xlsx contains the projected times for the 5,524 runners who did not finish in 2013 and for whom we produced projections. All seven methods were used, including the BAA ("Constant Pace") projections and Raymond Britt's method.

The file anovacode.txt contains the source code for the ANOVA analysis in R.

The file rescaledKNNcode.txt contains the source code for the rescaled KNN analysis in R. This code was used for all the analyses in the "Looking Forwards ... " section including Figures 7–9.

The file BM_knn.m contains the source code for the KNN analysis in Matlab format. This code was used to create the KNN results presented in the "Results" section and is described in the "Methods" section.

The files BM_lm.m and Pred_lm.m contain the source code for the linear regression analysis in Matlab format. This code was used to create the linear regression results presented in the "Results" section and is described in the "Methods" section.

SVDFold.py is Python code for the SVD analysis.

Other source codes will be posted later.