New Haven Road Race Analysis

Pei Tao

Abstract

My data is from the New Haven Road Race, 20k, from 2013-2017. I included Name, Age, Gender, Year, Division, and Nettime for each run from the original scrape of the website. Over time, I added variables such as temperature and humidity of the running day, along with calculating differences between the times a person got between different years and the difference in temperature and humidity on those days.

While some people only ran the race once, many people did the same race year after year, with varying results. Since these people are so consistant in running this race, I was curious if there was imporvement in their times over time, or even if a race could be partially predicted from information such as previous times, age, sex, and weather.

Overall, it did not seem like more races a person had under his or her belt helped them become faster, especially as they got older. Instead, weather conditions seemed more important in if a person would do better or worse.

Data

First, we collected “Event”, “Year”, “Name”, “Age”, “Sex”, “Div”, and “Nettime” from the results website. While some years at the Name in the correct format of “First Last” name, some years separted names in “First Name” and “Last Name”, which then I had to combine. In some years, there were no age given for the competitors and while some of them could be filled in by comparing these competitiors to those in years where age was collected and adding or subtracting the right number of years, others could not be. This was similar to Sex, such that some years collected the sex in a seperate column of the results table while other years did not. However sex cold be found in the Div column, and the general age range could be found there as well. Next, I cleaned out the runners who had the same names as other runners. This was doable as there were very few in the data set, but it was not possible to seperate them after running the age replacements. Then I calculated the difference between each runners time from that year versus the previous year, and the differences in temperature and humidity of that year and the previous year.

Analysis

```
repeatRunners <- read.csv("repeatRunners_withDiff.csv", as.is=TRUE)

lm.5 <- lm(LogDiff ~ + RaceNum + Div + LastYearTime + Temp + TempDiff + Humidity + HumidDiff,
           data=repeatRunners)
summary(lm.5)
```

```
##
## Call:
## lm(formula = LogDiff ~ +RaceNum + Div + LastYearTime + Temp +
##     TempDiff + Humidity + HumidDiff, data = repeatRunners)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59053 -0.04123 -0.00322  0.03922  0.59411
##
```

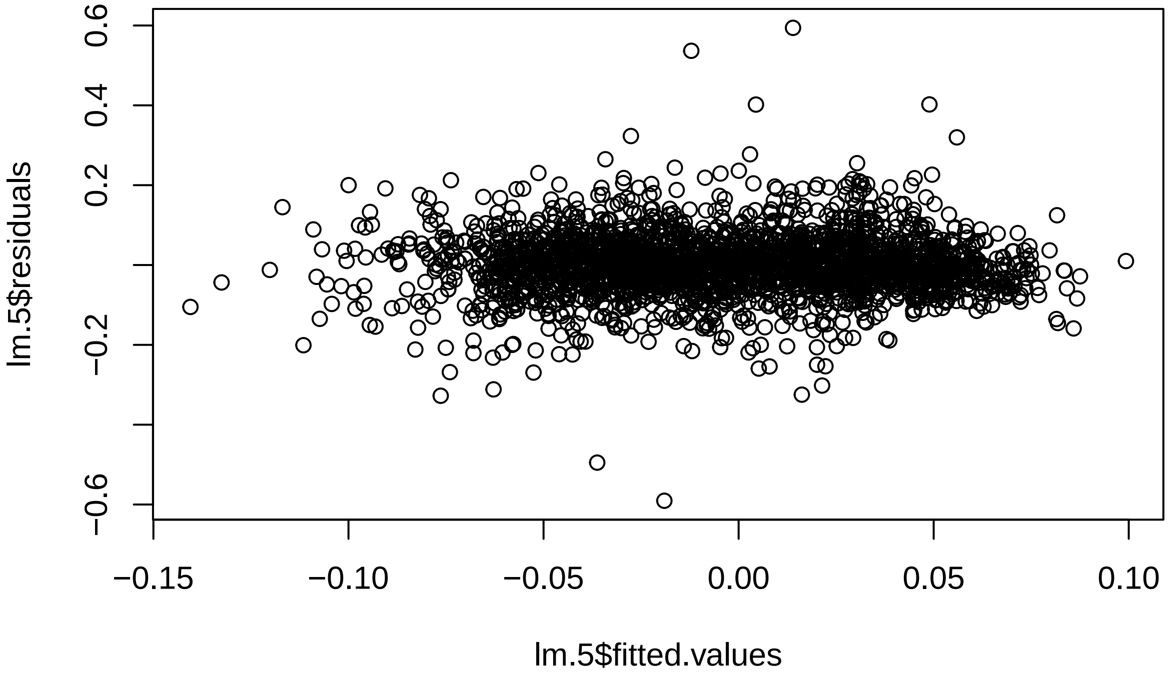
```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.338e-02  6.935e-02   1.346 0.178271
## RaceNum       3.769e-03  2.818e-03   1.338 0.181094
## Div13-19      3.040e-02  3.058e-02   0.994 0.320246
## Div20-29      2.673e-02  2.665e-02   1.003 0.315889
## Div30-39      3.267e-02  2.647e-02   1.234 0.217250
## Div40-44      4.573e-02  2.660e-02   1.719 0.085750 .
## Div45-49      3.818e-02  2.655e-02   1.438 0.150545
## Div50-54      5.076e-02  2.658e-02   1.910 0.056260 .
## Div55-59      5.525e-02  2.665e-02   2.073 0.038262 *
## Div60-64      7.234e-02  2.677e-02   2.702 0.006941 **
## Div65-69      7.413e-02  2.734e-02   2.712 0.006744 **
## Div70-74      9.450e-02  2.849e-02   3.317 0.000923 ***
## Div75-79      1.188e-01  3.828e-02   3.104 0.001932 **
## DivOVRAL      3.243e-03  3.115e-02   0.104 0.917080
## LastYearTime -9.224e-04  8.121e-05 -11.359  < 2e-16 ***
## Temp         -2.271e-03  1.157e-03  -1.964 0.049696 *
## TempDiff      5.658e-03  5.726e-04   9.881  < 2e-16 ***
## Humidity      1.746e-03  4.073e-04   4.287 1.88e-05 ***
## HumidDiff    -2.856e-04  2.692e-04  -1.061 0.288904
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07803 on 2437 degrees of freedom
## Multiple R-squared:  0.2046, Adjusted R-squared:  0.1988
## F-statistic: 34.83 on 18 and 2437 DF,  p-value: < 2.2e-16
```

I ran a lot of different linear models that all told more or less the same things, but at times also had various differences, such as in one linear model, it seemed like the number of races a person had run by that point was significant in how much it increases race times, which could be considered an interested find, but while this model does not point that out, it does say some other interesting (and potentially similar) things.

First, the older divisions, from Div60-64 to Div75-79, were significant in how being in such a division increases race time between years. This is understandable as once someone hits that age, they are far from the prime and with each passing year, regardless of how much training they do, their bady probably deteriorates more than can be compensated for.

Temperature was also a siginificant factor in how well someone did compared to the previous year. The New Haven Road Race happens during Labor Day weekend, in early September, where temperatures in New Haven are generally nice, as it's late summer, and it's very very rare to have a very cold day. However, there is still the chance of a siginifically warm day, with temperatures reaching above the 80s, where it's almost universal to consider such weather uncomfortable to move in. As we can see, the TempDiff is significant, where the high the temperature difference, the slower the time that year is.

```r
plot(lm.5$residuals ~lm.5$fitted.values)
```



Code Appendix

```
# fiveYears is the original scrape from the New Haven website
# (with some corrections such as in the names)
data <- read.csv("fiveYears.csv", as.is=TRUE)
uniqueNames <- unique(data$Name)

# we're going to make a list of the runners who have ran in more than one race
repeatRunners <- data.frame(Event=as.Date(character()),
                            Year=integer(),
                            Name=character(),
                            Age=integer(),
                            Sex=character(),
                            Div=character(),
                            Nettime=integer())

# this adds the runners to the dataframe above
# for runners who do not have unique names and shows up more than once per year, we delete those
for (ii in 1:length(uniqueNames)){
  name <- uniqueNames[ii]
  temp <- data[which(data$Name == name),]
  if (nrow(temp) > 1){
    years <- unique(temp$Year)
    if (length(years) == nrow(temp)){
      repeatRunners <- rbind(repeatRunners,temp)
    }
  }
}
```

```
# temperature and humidity for each year we might consider
temps <- data.frame(Year=c(2017, 2016, 2015, 2014, 2013, 2012),
                    Temp=c(68, 70, 72, 80, 74, 70),
                    Humidity=c(73, 51, 74, 82, 92, 79))

# list of unique names, aka each person we are considering
uniqueNames <- unique(repeatRunners$Name)

for (i in 1:length(uniqueNames)){
  name <- uniqueNames[i]
  rows <- which(repeatRunners$Name == name)
  for (j in 1:(length(rows)-1)){ # so for each of the years of each runner, but not the first year they
    # difference
    repeatRunners$Diff[rows[j]] <- repeatRunners$Nettime[rows[j]] -
      repeatRunners$Nettime[rows[j+1]]
    # run time last year
    repeatRunners$LastYearTime[rows[j]] <- repeatRunners$Nettime[rows[min(j+1, length(rows))]]
    # number of races run by that year
    repeatRunners$RaceNum[rows[j]] <- length(rows) - j + 1
    # log diff
    repeatRunners$LogDiff[rows[j]] <- log(repeatRunners$Nettime[rows[j]]) -
      log(repeatRunners$Nettime[rows[j+1]])

    year <- repeatRunners$Year[rows[j]]
    # temperature
    repeatRunners$Temp[rows[j]] <- temps$Temp[which(temps$Year == year)]
    # humidity
    repeatRunners$Humidity[rows[j]] <- temps$Humidity[which(temps$Year == year)]
    # temp diff from previous year ran
    repeatRunners$TempDiff[rows[j]] <- repeatRunners$Temp[rows[j]] -
      temps$Temp[which(temps$Year == repeatRunners$Year[rows[j]+1])]
    # humidity diff from previous year ran
    repeatRunners$HumidDiff[rows[j]] <- repeatRunners$Humidity[rows[j]] -
      temps$Humidity[which(temps$Year == repeatRunners$Year[rows[j]+1])]
  }
  repeatRunners <- repeatRunners[-rows[length(rows)],]
}

# combining the divisions so they aren't seperated into Male and Female
# figured this would be helpful in reducing the number of divisions
# and because male and female can be considered under sex
repeatRunners$Div <- gsub("M", "", repeatRunners$Div)
repeatRunners$Div <- gsub("F", "", repeatRunners$Div)
```