New Haven Road Race Analysis

Yan Qi

Abstract

In this paper, my primary goal is to build a model to predict the reasonable performances for each repeat runner in 5k New Haven Race, construct Predict Interval to help quantify any significant improvement or regress over time on an individual level. The data used for this analysis was scraped from the New Haven Road Race website, a database containing participants' basic features such as age, sex and performance.

Some descriptive statistic plots were used to get intuitions of the data, and simple linear regression was constructed on centered performances (one's Nettime minus his/her performance last year) to confirm that both age and sex have significant predictive power. While multi-axis plot cannot be used as guidance at most of time, weather on the race days do have some effect on runners' performance. Nonetheless, there are still more unknown factors between different years, that brings up significant different performance distributions among years.

Considering the huge range of age and the pattern for our data, I find that one of the most reasonable model is a Random Effect Model, which divides random error into the one within one person's data and the one across different individuals, and give us narrower prediction intervals.

In the light of the random effect model, we can claim that male runners generally doing better/regress slower than female runners, and people run slower when they grow older.

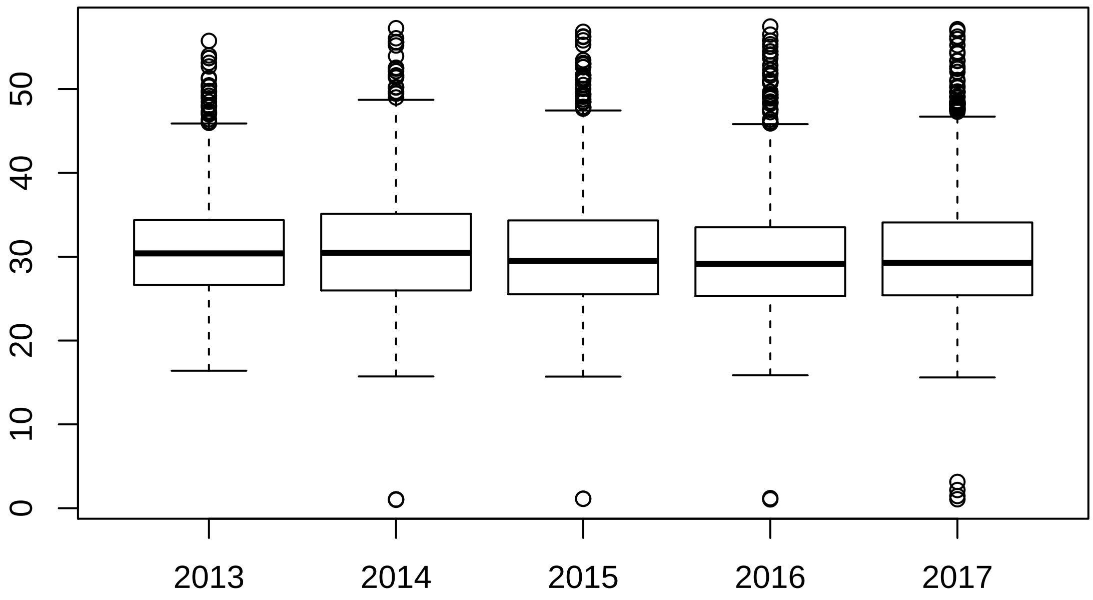
At last, I applied the Random Effect Model to the centered data, and confirmed that the across-individuals error term is zero. That means both the simple LS constructed on centered data and the Random Effect Model constructed on original data are reasonable to use to quantify individuals' improvements. The p-values calculated by those models also have consistent shapes of distribution.

Data

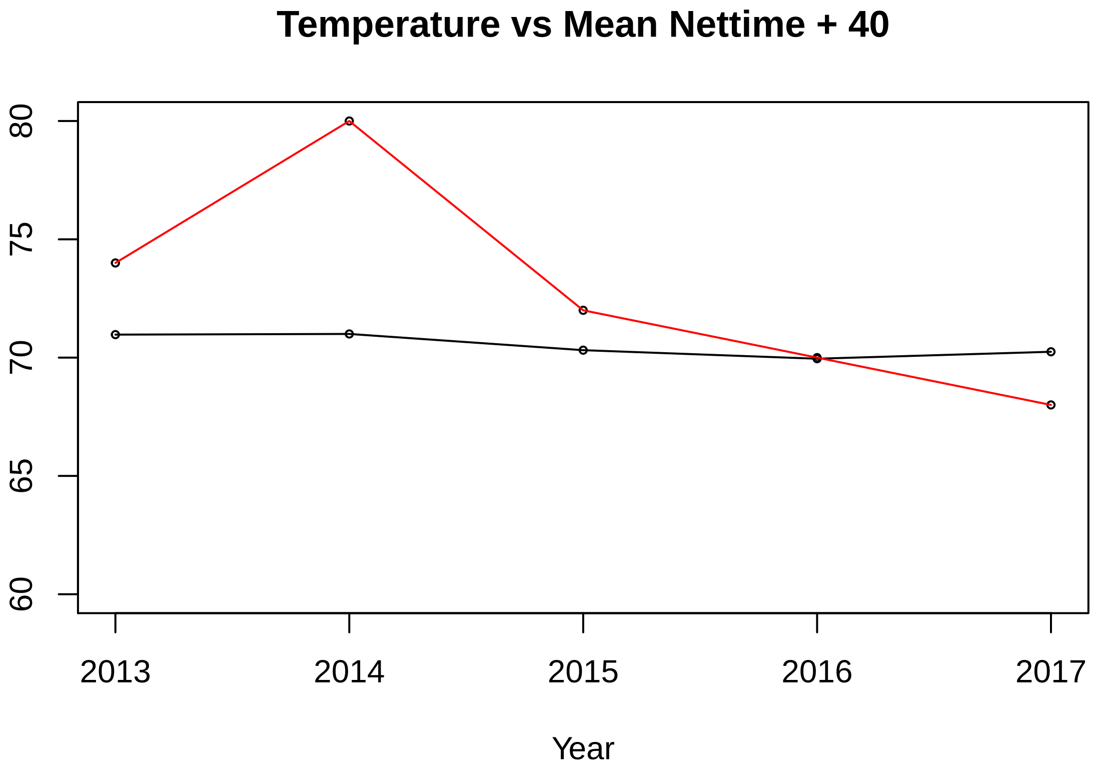
Before any modeling, I firstly removed all "Runner Unknown", then took step to find repeat runners who at least appear twice in the 5-year data. By checking if one name appears more than once in the same year, I found 41 people with the duplication name problem. As there are only 150 data related, I tossed them out. Using the cleaned data, I was able to fill in missing values for Age in 2016 and 2017, for those who also participated the race before 2016. Missing or inconsistent values of Sex were also addressed based on Name. Two new variables were added. One is the temperature on the days of races, the other is the "Dif" variable, the differences of Nettime between 2 successive races for each repeat runner. There were a few more individual outliers with absolute value of Dif larger than 30, I removed them too.

Analysis

Variable Selection



In order to gain a preliminary sense of which variable of the data should be included in the model, I looked at the plot of means and a boxplot of Nettime of difference years. The median of each year is rather stable, while the mean is influenced by extreme values. When I plot the shifted mean Nettime and Temperatures on race days in the same graph, they seem to have some relation. But unstandardized plot may be misleading sometimes.

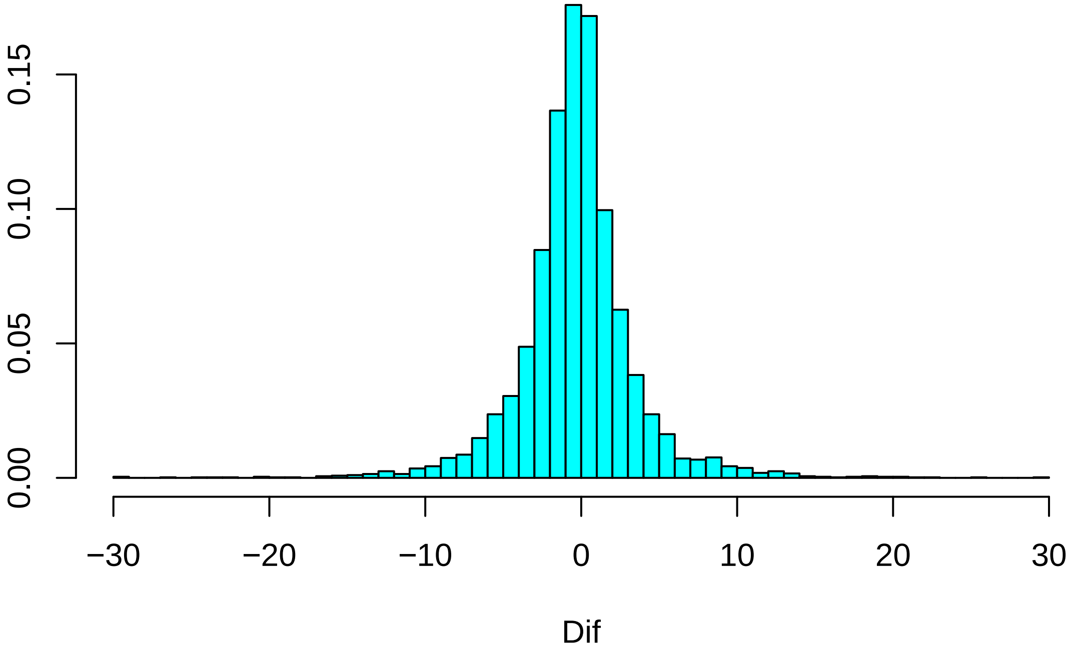


So, I carried out a simple linear regression of the Dif with regard to all the variables we have: Age, Sex, Year (as factor) and Tem.

Temperature on the race days do have some effect on runners' performance, but not much, with a negative coefficient of 0.024 and a p-value of 0.069. Nonetheless, there are still more unknown factors between different years that brings up significant different performance distributions among years. Besides, keeping both Tem and Year in the model brings up the multicollinearity problem, and there is no improvement of predictive power of our model, in both cases R-squared is 0.3329. The reason is that as we only have 4 data for temperatures, it just acts similar as a categorical variable. And all the information carried by Tem could be included in Year. Thus, I decided to remove the Tem variable.

Simple Linear Regression

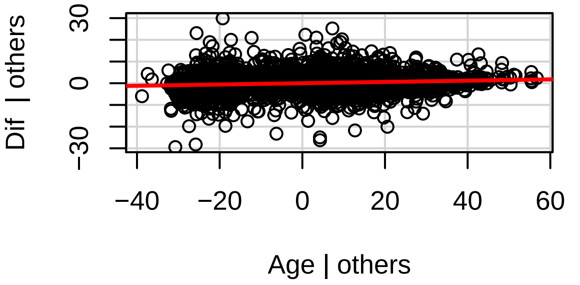
So far, using the basic analysis as guidance, I decide to include Sex, Age, Year (as factor) in my model of predicting Dif. In this LS model, assumption is that for each runner, he has his own mean Nettime, but the changes across years are all follow a same distribution with $\mu = 0$ and same standard deviation. The distribution of Dif is plotted in the following histogram, which has a bell shape.

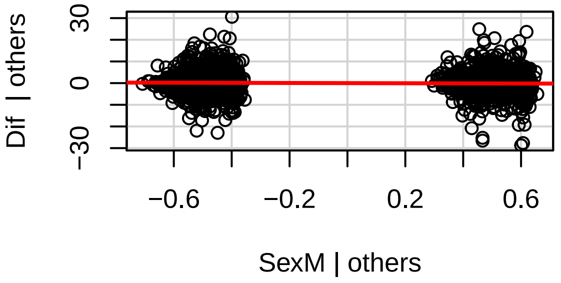


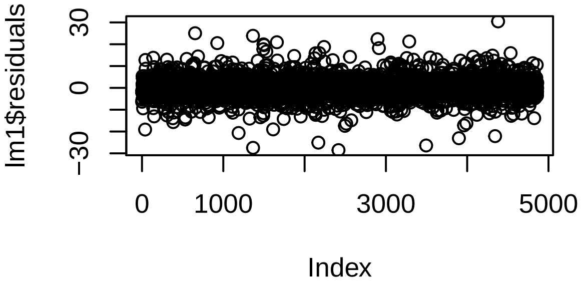
All variables are significant, but still cannot explain much of the variance. R-square is only 0.03329. The regression line has very small slopes and very bad predictions. One possible explanation of the poor fitting is that our data does not meet the independent requirement between each other. While I think the main issue here is the lack of data.

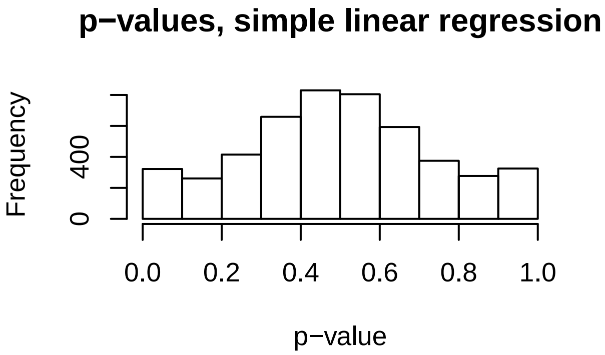
Added value plots give us an intuition of how much of the fluctuation of Dif is explained by our variables, I choose 2 of them to display below. See Appendix for the full summary of the model.

After running the regression, I also visualized the residuals. It has a large range because of the poor prediction, but looks relatively well-behaved with no obvious trends.





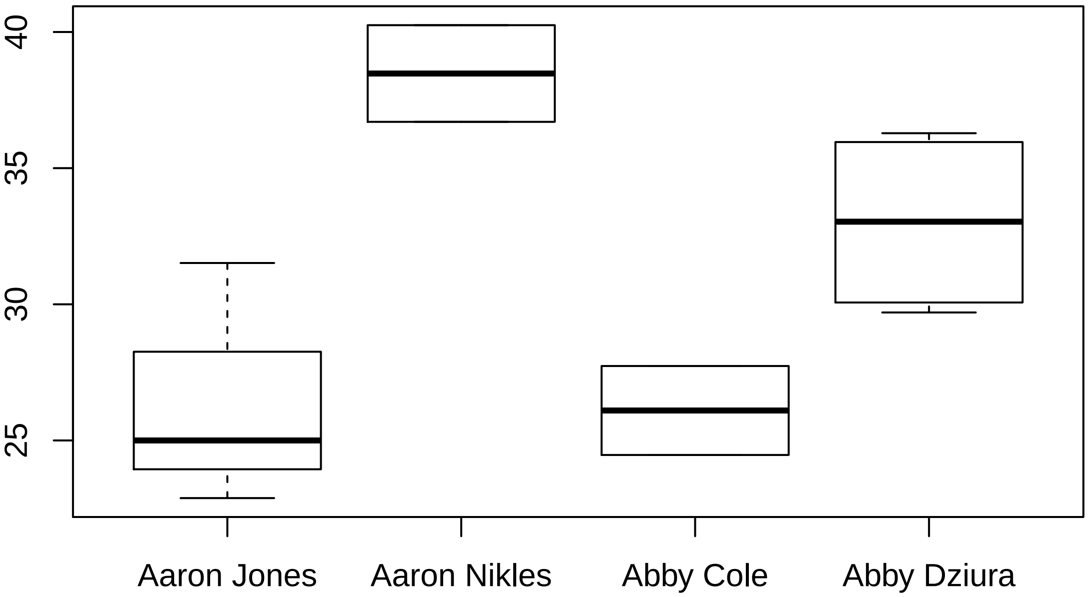




The model implies that male runner generally doing better/regress slower than female runner, and people run slower when they grow older. Due to our limited data, this model tells us little about why people tend to run faster from 2014 to 2015, 2015 to 2016, but slower from 2016 to 2017. I calculated the one-tail p-value for each runners' Dif base on the Prediction Interval given by the simple linear regression model. The standard deviation used here was 3.704 (the one for prediction intervals).

Random Effect Model

Although the result above seems reasonable considering our lack of data, I realized that there is considerable variation among runners, because of the huge range of ages. The variation among people cannot be explained well with our current dataset. But variations within one person's data should be much smaller. If we can find it, our prediction interval and p-value would be more accurate, giving us more confidence to determine whether an individual improved or not. So, I visualized the data grouped by Name, for example, the first 4 repeat runners have performances as below. A random effect model would be suitable to help us split the variation within one person's performances and the random effect among people.

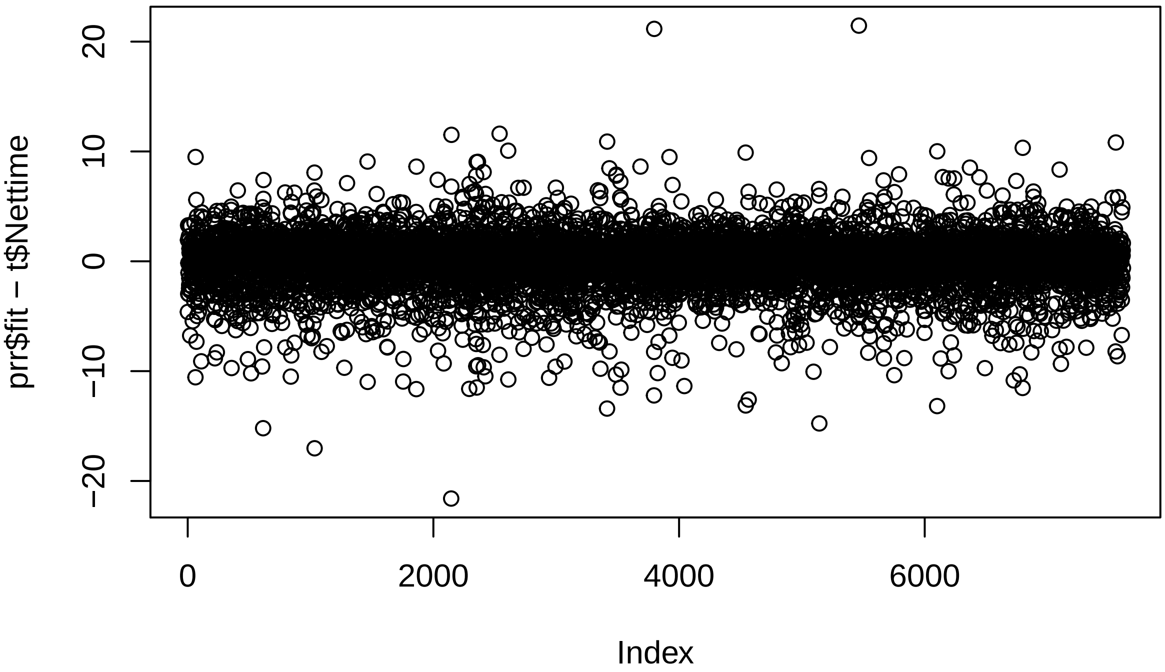


The assumption here is: each person has a mean performance, which is predicted by fixed effects (Sex, Age, Year). There is noise between different measures for one person, and random effect among different people.

$$y_{ij} = mu + a_i + e_{ij}$$

I constructed the regression model for the original data Nettime. Fixed effects are still significant. We have positive coefficient for Age and negative coefficient for SexM just as before, but in different scales.

The standard deviation within each group is 2.870, which is smaller than 3.703, given by the simple linear regression model. And the standard deviation among groups is 5.474. Besides, the prediction was also improved according to the plot of residuals. There is no significant trend in the residuals and the range shrank.



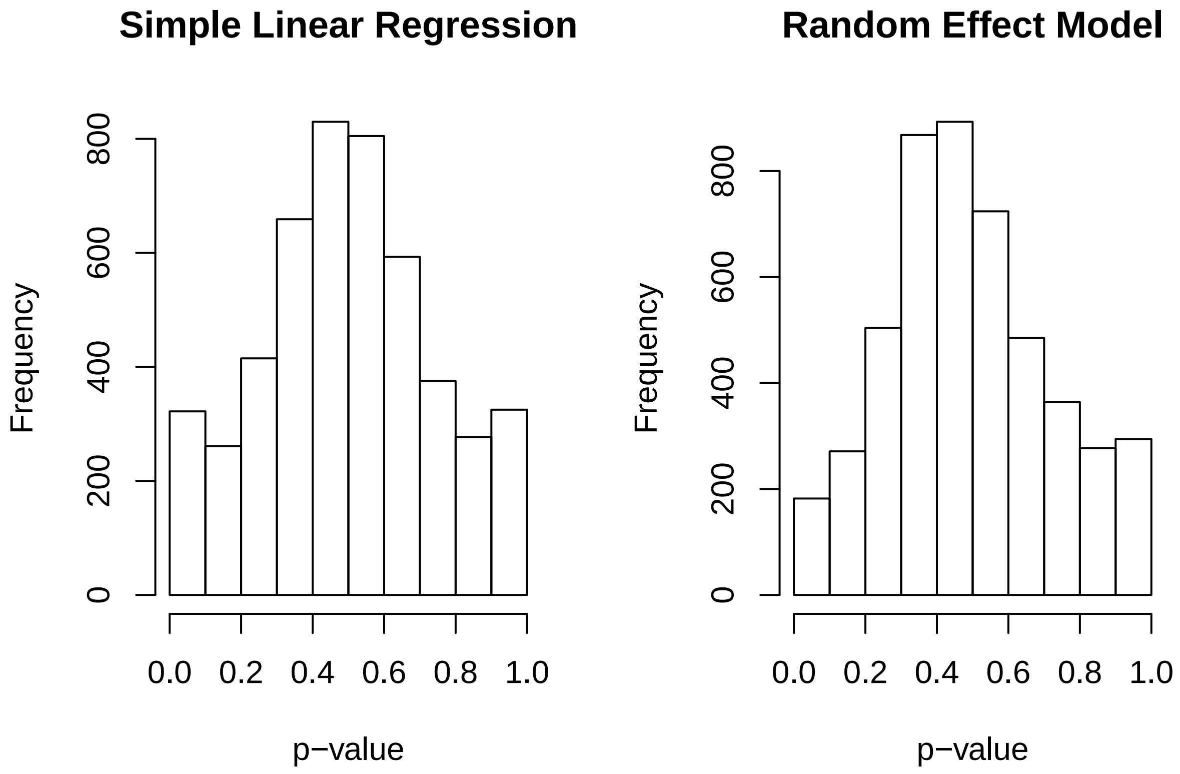
I use this model to calculate one-tail p-value for runners, part of the results including fitted values and p-values are shown below.

```
##             Name Year  Nettime    fitrr       rrp
## 3   Aaron Jones 2013 31.51667 26.90535 0.9459444
## 4   Aaron Jones 2014 25.00000 26.93022 0.2506148
## 2   Aaron Jones 2015 22.88333 26.15723 0.1269894
## 5  Aaron Nikles 2014 40.25000 37.28966 0.8488430
## 6  Aaron Nikles 2015 36.70000 36.54038 0.5221764
## 9     Abby Cole 2015 27.73333 26.68718 0.6422632
## 10    Abby Cole 2016 24.46667 26.32202 0.2589882
## 12  Abby Dziura 2013 36.28333 33.29692 0.8509621
## 14  Abby Dziura 2015 29.70000 32.95737 0.1281919
## 13  Abby Dziura 2016 30.43333 32.71581 0.2132211
```

Model Comparision and Conclusion

To compare the performance of the two models, I applied the random effect model to the Dif variable, against same set of independent variables. The variance across different groups turned out to be 0, which confirms the assumption of the distribution of Dif at the beginning of this part.

To conclude, both models seem reasonable to quantify runners' improvement/regress, while the random effect model has a little bit more accurate prediction power. Actually, they do give us similar p-values, and the histograms of p-values from two models have similar shape. Except that the distribution of simple linear regression's p-values is kind of fat-tail due to the poor prediction of that model.



The coefficients in random effect model for Age and Sex is 0.1039 and -4.5718 respectively. As both model are consistent, we can claim that male runners generally doing better/regress slower than female runners, and people run slower when they grow older with some confidence.