Exploratory Data Analysis Report

Athletes will dedicate several years of training into their performance at the Olympics. The goal is to push the limits of the human body in athletic excellence. We will be focusing on track events that involve strictly running, excluding the marathon. Some of these races are finished in under 11 seconds and others last over half an hour. How can we best predict an Olympic track athlete’s finishing time in these races? What is the relationship between an athlete’s finishing time, adjusted for distance, at the Olympics and their sex, age, weight, height, year of competition, and the athlete’s nationality, their country GDP, and population? Our goal is to model an athlete’s race time with athlete-level characteristics and their team’s country-level characteristics.

While investigating relationships between these variables, we kept in mind that we were headed toward a hierarchical model and so, in an organizational strategy, we analyzed the variables by their levels. First, we will explore variables at level 1, which is an athlete’s race, then level 2, the country the athlete is from, and then cross-level relationships. A level 1 unit is athlete’s race that is defined by the athlete, with characteristics such as age, height, weight, and sex, and defined by the race, with characteristics such a distance and year.

Looking at the correlation plot and corresponding table of values of level one variables and the response, there are some obvious pairs of variables with strong relationships. Height and weight are highly positively correlated (0.824). Distance of the race and the finishing time are nearly perfectly correlated (0.998). Then, there are some more interesting, albeit weaker, relationships. Both weight (-0.41) and height (-0.34) are negatively correlated with finishing time. Athletes that are taller or weigh more, presumably weight due to muscle, tend to have faster times. However, when we look at graphs of time versus height and time versus weight grouped by distance, we can see there are different slopes depending on the distance of the race. The effect of weight seems to matter less as the distance increases, while the effect of height peaks at the middle tier distance and matters less on especially short races or especially long races. Now, age is positively correlated with finishing, meaning the older an athlete the slower the time tends to result. Yet, when this is broken down by distance, there are positive slope for 800-meter races or longer and negative slopes for 400-meter races and shorter. Seemingly, older athletes perform better in longer distance races. Another nuance to age is that there is a positive correlation between year and age (0.1943), suggesting that athletes are competing longer into their careers. Remarkably year and finishing time has a very low (positive) correlation, but the plots of time versus year grouped by distance does show a downward trend in every distance and the coefficient for year in the regression line predicting time (seconds) from year and adjusting for distance is very statistically significant (t-value = -5.5).

On average, males tend to have lower finishing times than females. Maybe a less know difference between males and female athletes, across Olympic history there have been almost triple the amount of male medaling athletes compared to female (413 males; 172 females). Also, there are less competitors overall as the distance increases and males’ finishing times appear to vary more as distance increases while females’ finishing times variation seems to stay about the same. In short, at level 1 there seems to be graphical and numerical indicators to include interaction with age and distance, weight and distance, height and distance, age and year, and sex and distance.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | age | height | weight | timeSecs | year | dist |
| age | 1 | -0.090 | -0.114 | 0.238 | 0.194 | 0.238 |
| height |  | 1 | 0.824 | -0.337 | -0.153 | -0.327 |
| weight |  |  | 1 | -0.414 | -0.187 | -0.411 |
| timeSecs |  |  |  | 1 | 0.055 | 0.998 |
| year |  |  |  |  | 1 | 0.070 |
| dist |  |  |  |  |  | 1 |

