5-2021

Regression Analysis of Pacing When Running a Marathon

Hawkin Starke

Regression Analysis of Pacing When Running a Marathon

Hawkin Starke and Justin R Chimka

University of Arkansas, Fayetteville

Word Count: 2679

Abstract

Regression analysis can be an effective way of examining performance in the marathon event. By splitting up the race into segments or in runner terminology “splits” the significance of each segment as it relates to the total finish time can be explored. Because the idea of splits is already ingrained into the minds of runners, it makes intuitive sense to use these as the metrics to define a race. Additionally, marathons generally make participant age and gender date publicly available which can then be used to find trends within specific demographics. This tailors trends to smaller groups of people, making the lessons from these trends more easily applied during the marathon. The most popular warning within the marathon community is that of the “fast start” which is translated to mean running faster than your average pace at the beginning of the race and consequently slowing down through the remainder of the race. Because of this, after segregating our data into four subsets, each runner’s pace in the first 10 kilometers (23.66%) of the race was plotted against their total finish time. In three out of the four subsets of the data the anecdote appeared to be clearly substantiated as runners who started slowest in relation to their mean race pace tended to have lower total finish times, and finish time generally increased as the percentage above mean race pace of runners increased.

Keywords

Regression, Marathon, Multivariate Statistical Analysis

Introduction

The marathon is the defining endurance event across the globe, and millions of participants toe the start line of events each year hoping to best their former selves. For many, the measuring stick which they use to compare themselves is their finish time. However, many large marathons actually record the time that runners cross thresholds throughout the race in addition to their time at completion of the event. This is significant because even in the unlikely instance in which two runners finish a marathon in the exact same amount of time, it is probable that they achieved their "journey" was different i.e. their pacing through given segments of the race. The following analysis examines patterns in these "split times" suggesting that some segments of the race may be more significant than others in the determination of a runner's finish time. Trends that can be identified within these more significant splits can be used as a prescriptive tool for runners who are hoping to reach a personal best in an upcoming race. As an added benefit to runners, the model presented within this paper includes age and gender. This allows trends within subsets of participants to be identified, leading to a more reliable application of this research.

Literature Review

Pacing in the marathon event has been the subject of many studies within the last twenty years, as followers of the sport have attempted to leverage statistical analysis in order to optimize performance. Papers written by Keogh, et al. (2020); Haney (2010); de Koning, et al. (2011); and Abbiss and Laursen (2008) all approach the topic from varied and substantive perspectives. The subtopics surrounding the athletic phenomenon are wide examining possible influencers, most notably age and gender. Separate papers written by Buman, et al. (2008); March, et al. (2011);

and Trubee (2011) come to conclusions that are generally backed up by common anecdotal observations within the running community such as, women generally keep more even splits than men, and older runners keep more even splits than younger runners. Other papers like the one written by Summers, et al. (1982) focus only on middle aged, middle paced runners that make up the largest portion of the population examined in this paper. Less quantitative factors that can influence marathon performance have been investigated as well, such as emotions of runners throughout the race documented in the paper written by Baron, et al. (2011). There has also been a significant volume of research dedicated to the coaching of participants in the marathon and what strategies are most effective in sustaining improvement of performance (Denison, 2007). Papers published in the *International Journal of Performance Analysis in Sport* cover related subtopics in the world of running such as examining the popular run-walk marathon pacing strategy and the trends within pacing in a 24 hour ultramarathon. (Nolan et al. 2020 and Takayama et al. 2016, respectively).

Materials and Methods

The data used in this analysis were gathered from the Little Rock Marathon website where the results of each race are published online and made publicly available. We did not exclude any participants based on a time threshold, as we were interested in comparing the trends in performance across all fitness levels. Over two years (2016-2017) the Little Rock Marathon had 4129 complete participant records. The data were segregated by year in order to monitor any trends year over year and examine the possible effects of race day conditions such as weather. In 2016 and 2017 the Little Rock Marathon split its course into four segments:

- The starting line to the 10K timing mat
- The 10K timing mat to the half marathon timing mat
- The half marathon timing mat to the 21 mile timing mat
- The 21 mile timing mat to the finish line

The segments were aptly named TenK, Half, TwOne, and Final, respectively. In the interest of clearly illustrating the splits used in this work, a graphic of the Little Rock Marathon course map has been placed below (Timing Mats were marked using a black oval; these timing mats represent the beginning of the upcoming split and the end of the previous one).

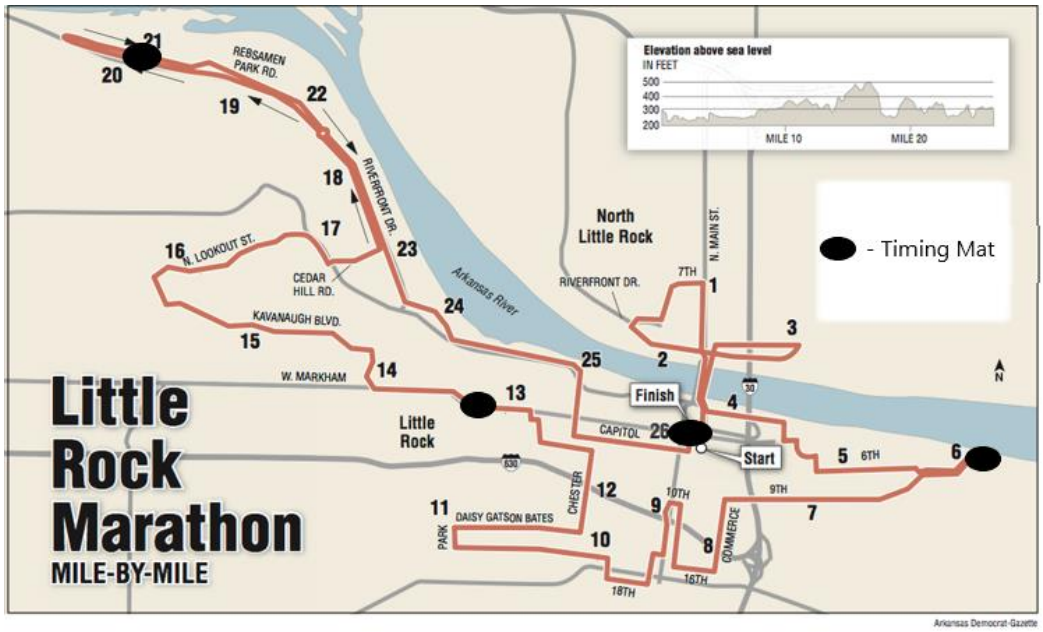


Figure 1. Little Rock Marathon Course Map

We removed one of the time splits, so the model could not entirely determine finish time (if we included all four splits a regression model would simply sum the four splits to "predict" finish time, therefore not predicting anything at all). Final (the segment from the 21 Mile timing mat to

the finish line) is the segment that was removed from consideration. Multiple linear regression using five factors was applied to the finish times of each year. The five factors were Age, Gender, 10K Pace, Half Marathon Pace, and 21 Mile Pace (the three variables representing 10K Pace, Half Marathon Pace, and 21 Mile Pace are referred to as TenK, Half, and TwOne within the regression models).

Results

We begin our analysis by fitting main effects models to each dataset (2016 and 2017). These models take as the response Finish Time and independent variables Age, Gender, 10K Pace, Half Marathon Pace, and 21 Mile Pace.

- In the 2016 dataset insignificant variable 21 Mile Pace significantly interacts with 10K Pace, so it was used to separate the 2016 dataset into two subsets, where 10K Pace becomes significant for slower 21 Mile Pace runners, and can be dropped from the model of faster 21 Mile Pace runners.
- In the 2017 dataset insignificant variable 10K Pace significantly interacts with Age, so it was used to separate the 2017 dataset into two subsets. Even so Age is ultimately dropped from both models, slower 10K Pace runners and faster 10K Pace runners.

These decisions are illustrated in Figure 2 (to clarify, TwOne High corresponds to the slower group of runners and TwOne Low corresponds to the faster group of runners with the same logic applying to TenK High and TenK Low).Variables remaining in the four statistical models are all significant; coefficients and adjusted R-squared values are available in Table 1.

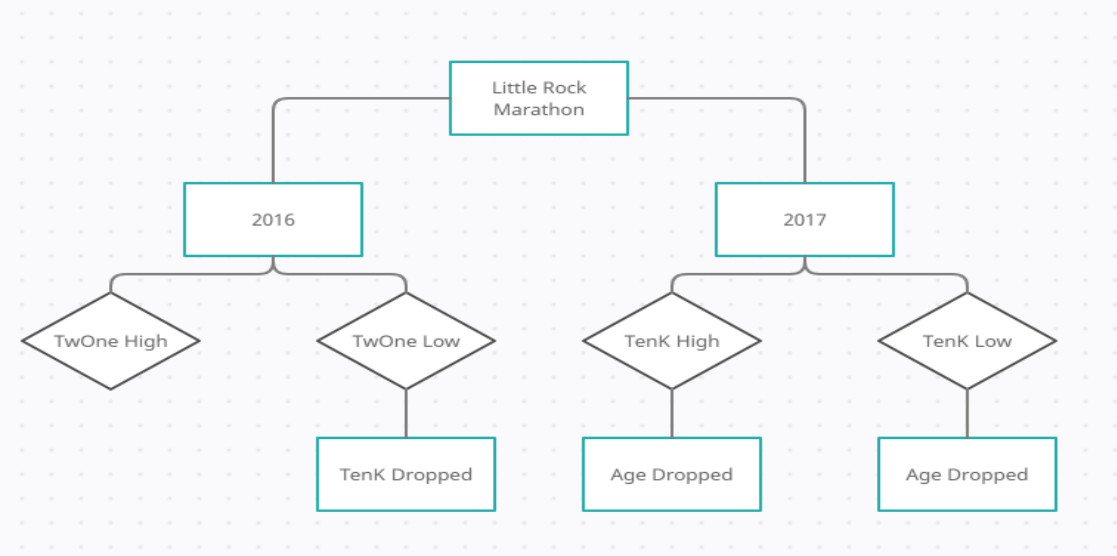


Figure 2. Tree-based variable selection

Table 1. Coefficients and adjusted R-squared of the four statistical models

Coefficients	2016 TwOne High	2016 TwOne Low	2017 TenK High	2017 TenK Low
Age	-0.00011	-0.0000065		
Gender	0.002494	0.002829	0.001594	0.000569
TenK	0.4954			
Half	2.9334	3.7742	11.900	10.878
TwOne			12.425	13.162
Constant	0.04936	0.01193	0.01802	0.016025
Adjusted R-Squared	0.8878	0.8853	0.9719	0.9822

Gender and the Half split are significant in all four of the models. In 2016, the magnitude of the T-value is much greater for the Half split than any of the other variables, making it the

most influential factor, controlling for other variables in the model. In 2017, the most influential factor is the TwOne Mile split (which was used to separate the data in 2016), closely followed by the Half split. These results support the observation that the Half split is the most important factor across the four models and a participant's performance during this split is a strong indicator of likely finish times for each participant. The adjusted R-Squared for the two 2017 models is higher than that of the 2016 models. One external factor is likely to be responsible for the inconsistency across years: Weather. In 2016 the race had a high temperature of 75 degrees Fahrenheit and a maximum sustained wind speed of 18.30 mph. Both of these measurements are well above optimal for distance running and therefore might have contributed to the unpredictability of times. This is especially true when compared to the conditions in 2017, where the high temperature was 62 degrees Fahrenheit and the maximum sustained wind speed was 16.11 mph. These measurements are much closer to desired conditions for performance and may contribute to other differences between years in our analysis.

After partitioning the two years of marathon data into four more meaningful subsets, we leveraged the data to serve as a predictive tool for runners who are looking for an improved methodological approach to the marathon event. The following graphs examine the relationship between %RP, a metric developed by Smyth (2018), and the finish time of participants. The equation for %RP (percent race pace) is shown below.

$$\%RP(Segment) = 100x\left(\frac{MRP - Pace(Segment)}{MRP}\right)$$

(MRP is an acronym for mean race pace, the average time per mile over the course of all 26.2 miles)

Each value on the x-axis represents multiple participants, because %RP values have been rounded to the nearest whole percent. Therefore the values for finish times are an average of all finish times belonging to participants with the corresponding %RP value. The "segment" from the above equation that we chose to examine is the first ten kilometers of the race. The reasoning behind this is anecdotal, as there have been many accounts of "starting too fast" and therefore depleting energy at such a rapid rate that a significant slowdown later in the race is inevitable. As evidenced by Figures 3 and 4 shown below it does seem that finish time tends to rise as %RP also rises. This is consistently and more dramatically the case in 2017, under better running conditions. In other words the faster a runner starts compared to their average race pace, the slower that runner tends to finish. Additionally, it is interesting to note how few participants run their first split more slowly than what turns out to be their average race pace. Another notable difference between the results of the 2016 and 2017 marathons is the gap between the slowest average finishing time for the fast group and the fastest average finishing time for the slow group. In 2017 the highest point on the fast group line approaches the lowest point on the slow group line while in 2016 there is an hour of finish time between these points. This creates two distinct groups of runners much more well defined than those in 2017. Again, a possible explanation relates to weather and the way in which it affects groups of runners differently. It appears that faster runners (generally better prepared) were able to mitigate the effects of heat much more effectively and in some cases avoid them entirely.

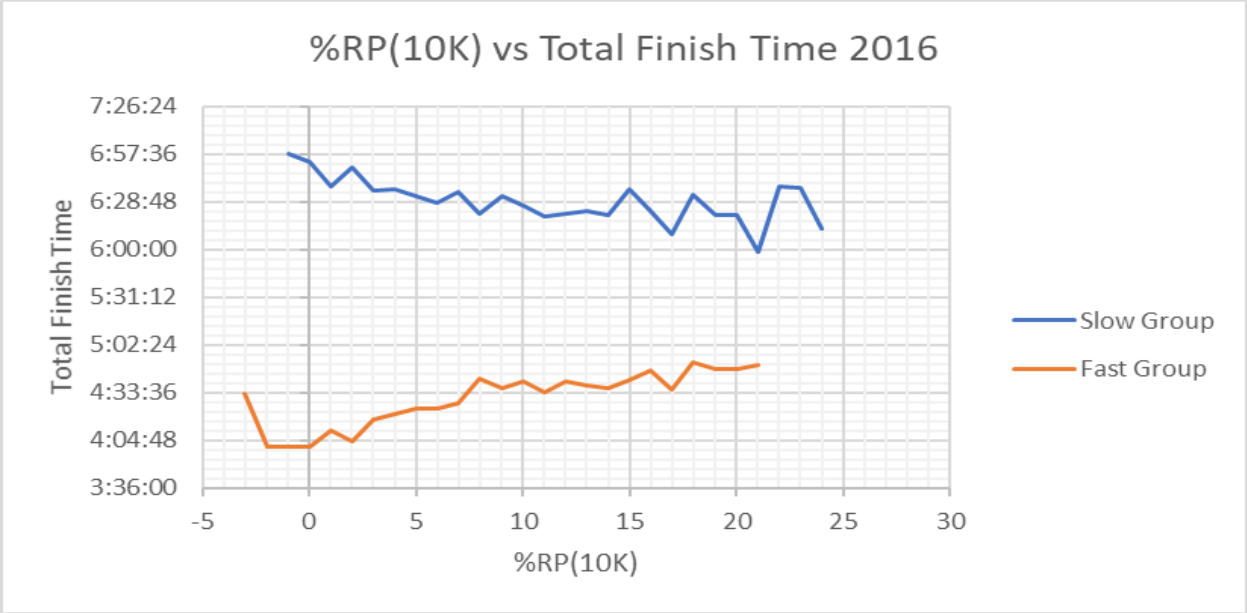


Figure 3. %RP(10K) vs Total Finish Time 2016

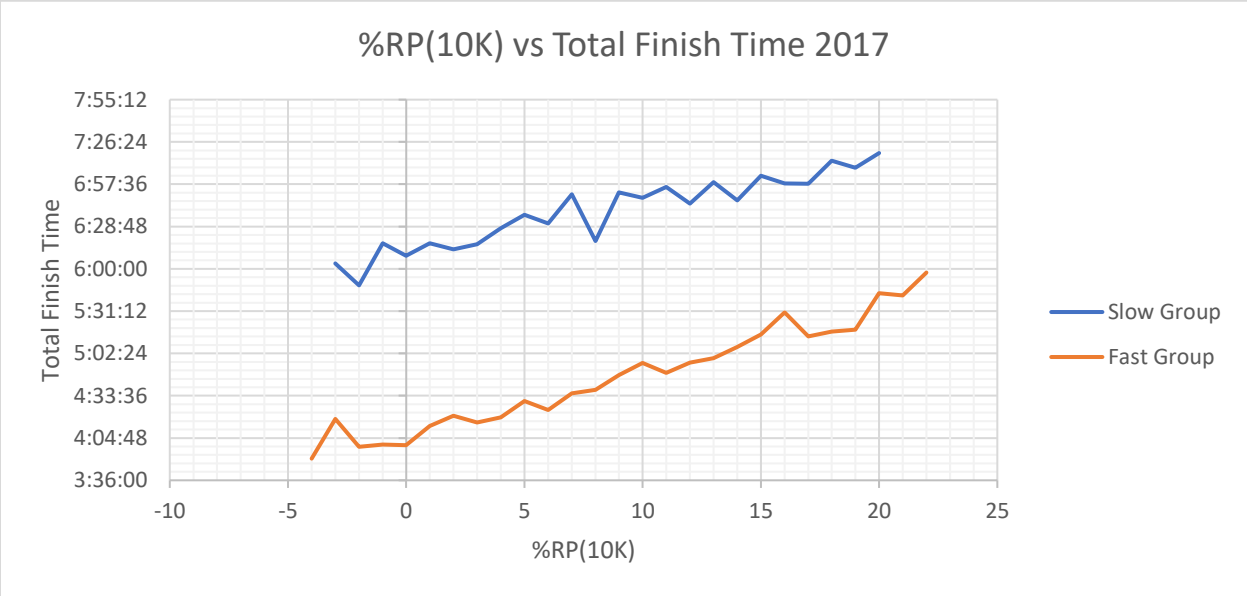


Figure 4. %RP(10K) vs Total Finish Time 2017

Discussion

Based on the results of our regression models, more than 80% of variation in finish time can be explained for participants in the marathon event given their age, gender, and their first three splits. Among them, the Half split, from the 10K timing mat to the half marathon mat is most consistently important to both fast and slow runners in both years. The TwOne Mile split from the half marathon timing mat to the 21 mile mat was also very important in that it was used to define fast and slow runners in 2016 (due to interaction among variables), and was the most important factor in determining 2017 finish times. Performance in 2017 seems consistently more predictable than 2016 according to adjusted R-Squared values, perhaps due to differences in weather conditions: 2017 had more favorable conditions for the Little Rock Marathon. Finally we have seen through %RP curves how finish times tends to rise as runners start the marathon at paces that cannot be maintained.

Future research could include model adequacy checking if there was interest in building confidence or prediction intervals for certain runners' finish times. A careful examination of model errors may lead an analyst to more theoretically appropriate statistical models that generate even more substantial results. Of course additional datasets from this and other marathons analyzed in a consistent manner may do more to verify the results we have seen here.

Acknowledgements

Hawkin Starke's research was supported by the Walton Foundation through the Honors Fellowship program at the University of Arkansas.

Declaration of Interest Statement

There were no conflicts of interest throughout the creation of the above manuscript.

References

Abbiss, Chris R, and Paul B Laursen. “Describing and Understanding Pacing Strategies during Athletic Competition.” *Sports Medicine*, vol. 38, no. 3, 2008, pp. 239–252., doi:10.2165/00007256-200838030-00004.

Baron, B., et al. “The Role of Emotions on Pacing Strategies and Performance in Middle and Long Duration Sport Events.” *British Journal of Sports Medicine*, vol. 45, no. 6, 2009, pp. 511–517., doi:10.1136/bjsm.2009.059964.

Buman, Matthew P., et al. “Hitting the Wall in the Marathon: Phenomenological Characteristics and Associations with Expectancy, Gender, and Running History.” *Psychology of Sport and Exercise*, vol. 9, no. 2, 2008, pp. 177–190., doi:10.1016/j.psychsport.2007.03.003.

De Koning, Jos J., et al. “Regulation of Pacing Strategy during Athletic Competition.” *PLoS ONE*, vol. 6, no. 1, 2011, doi:10.1371/journal.pone.0015863.

Denison, Jim. “Perspectives on Coaching Pace Skill in Distance Running: A Commentary.” *International Journal of Sports Science & Coaching*, vol. 2, no. 3, 2007, pp. 217–238., doi:10.1260/174795407782233128.

Haney, T.A. “Variability of Pacing in Marathon Distance Running.” *University of Nevada, Las Vegas*, 2010.

Keogh, Alison, et al. “The Determinants of Marathon Performance: An Observational Analysis of Anthropometric, Pre-Race and In-Race Variables.” *International Journal of Exercise Science*, vol. 13, no. 6, Aug. 2020, pp. 1132–1142.

March, Daniel S, et al. “Age, Sex, and Finish Time as Determinants of Pacing in the Marathon.” *Journal of Strength and Conditioning Research*, vol. 25, no. 2, 2011, pp. 386–391., doi:10.1519/jsc.0b013e3181bffd0f.

Nolan, William P., and Andrew R. Moore. “Run-Walk Marathon Pacing: the Energy Cost of Frequent Walk Breaks.” *International Journal of Performance Analysis in Sport*, vol. 21, no. 1, 2020, pp. 170–179., doi:10.1080/24748668.2020.1862493.

Smyth, Barry. “Fast Starters and Slow Finishers: A Large-Scale Data Analysis of Pacing at the Beginning and End of the Marathon for Recreational Runners.” *Journal of Sports Analytics*, vol. 4, no. 3, 2018, pp. 229–242., doi:10.3233/jsa-170205.

Summers, Jeffery J., et al. “Middle-Aged, Non-Elite Marathon Runners: A Profile.” *Perceptual and Motor Skills*, vol. 54, no. 3, 1982, pp. 963–969., doi:10.2466/pms.1982.54.3.963.

Takayama, Fuminori, et al. “Pacing Strategy in a 24-Hour Ultramarathon Race.” *International Journal of Performance Analysis in Sport*, vol. 16, no. 2, 2016, pp. 498–507., doi:10.1080/24748668.2016.11868904.

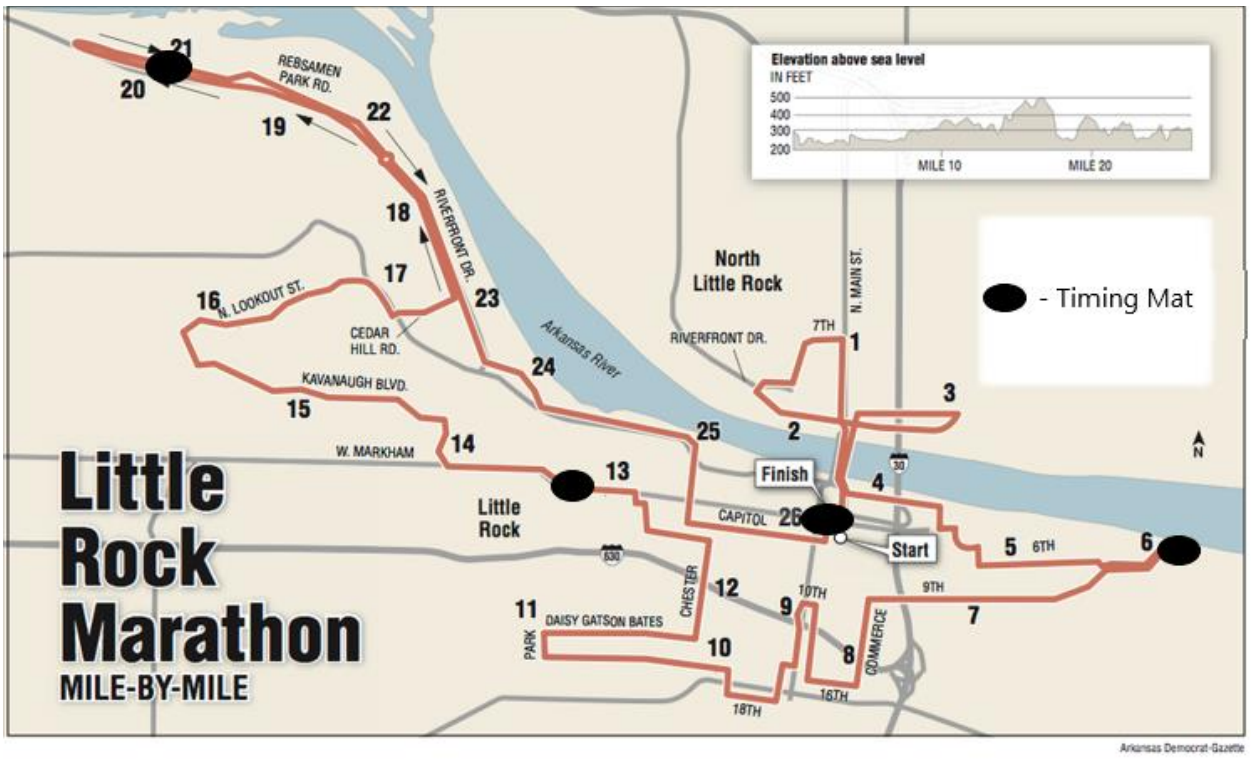
Trubee, N.W. “The Effects of Age, Sex, Heat Stress, and Finish Time on Pacing in the Marathon.” *University of Dayton*, 2011.

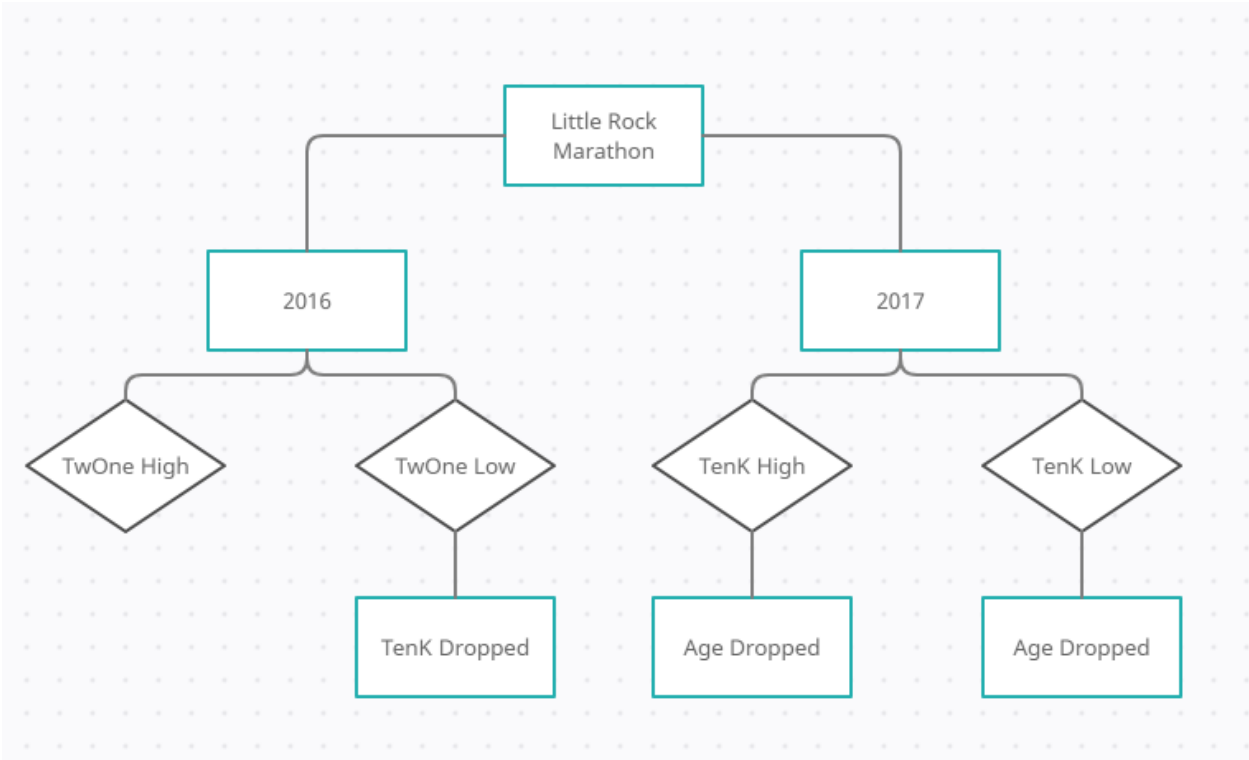
Tables

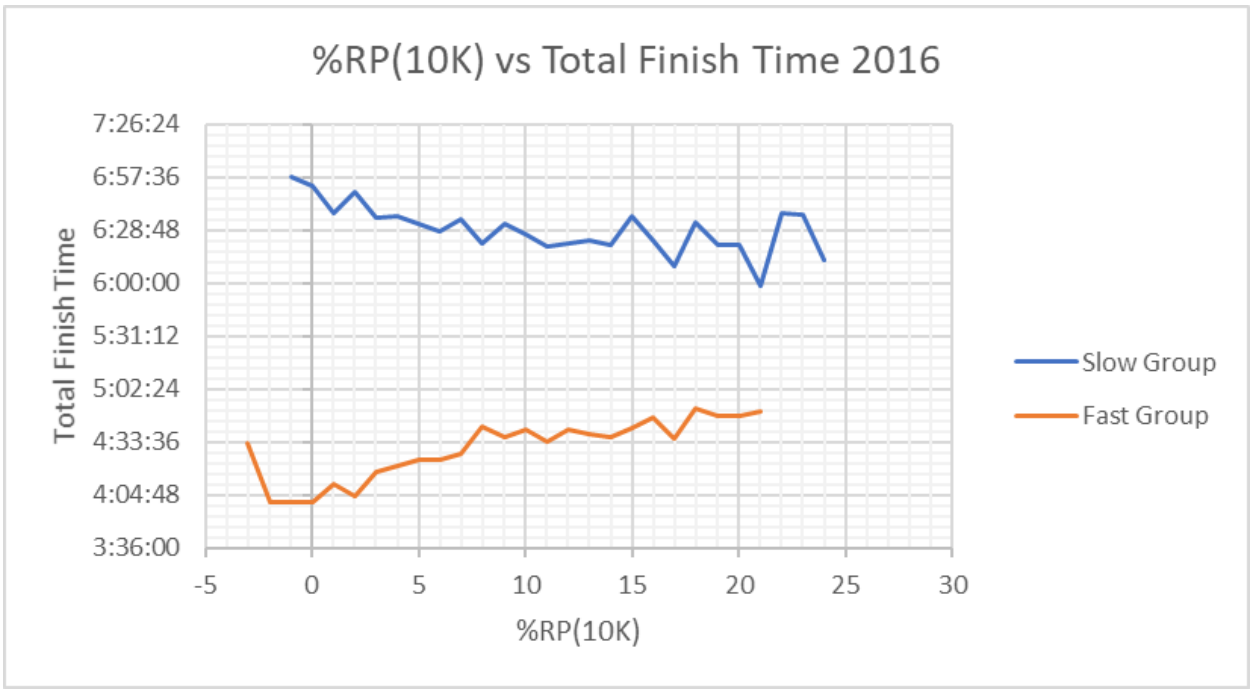
Table 1. Coefficients and Adjusted R-squared of the Four Statistical Models

Coefficients	2016 TwOne High	2016 TwOne Low	2017 TenK High	2017 TenK Low
Age	-0.00011	-0.0000065		
Gender	0.002494	0.002829	0.001594	0.000569
TenK	0.4954			
Half	2.9334	3.7742	11.900	10.878
TwOne			12.425	13.162
Constant	0.04936	0.01193	0.01802	0.016025
Adjusted R-Squared	0.8878	0.8853	0.9719	0.9822

Figures


Little Rock Marathon
MILE-BY-MILE
Elevation above sea level
IN FEET
500
400
300
200
MILE 10
MILE 20
- Timing Mat
North Little Rock
Little Rock
Start
Finish
REBSAMEN PARK RD.
RIVERFRONT DR.
N. LOOKOUT ST.
CEDAR HILL RD.
KAVANAUGH BLVD.
W. MARKHAM
Arkansas River
N. MAIN ST.
7TH
CAPITOL
9TH
10TH
16TH
18TH
CHESTER
COMMERCE
DAISY GATSON BATES PARK
Arkansas Democrat-Gazette


Little Rock Marathon
2016
2017
TwOne High
TwOne Low
TenK High
TenK Low
TenK Dropped
Age Dropped
Age Dropped


%RP(10K) vs Total Finish Time 2016
Total Finish Time
7:26:24
6:57:36
6:28:48
6:00:00
5:31:12
5:02:24
4:33:36
4:04:48
3:36:00
-5
0
5
10
15
20
25
30
%RP(10K)
Slow Group
Fast Group

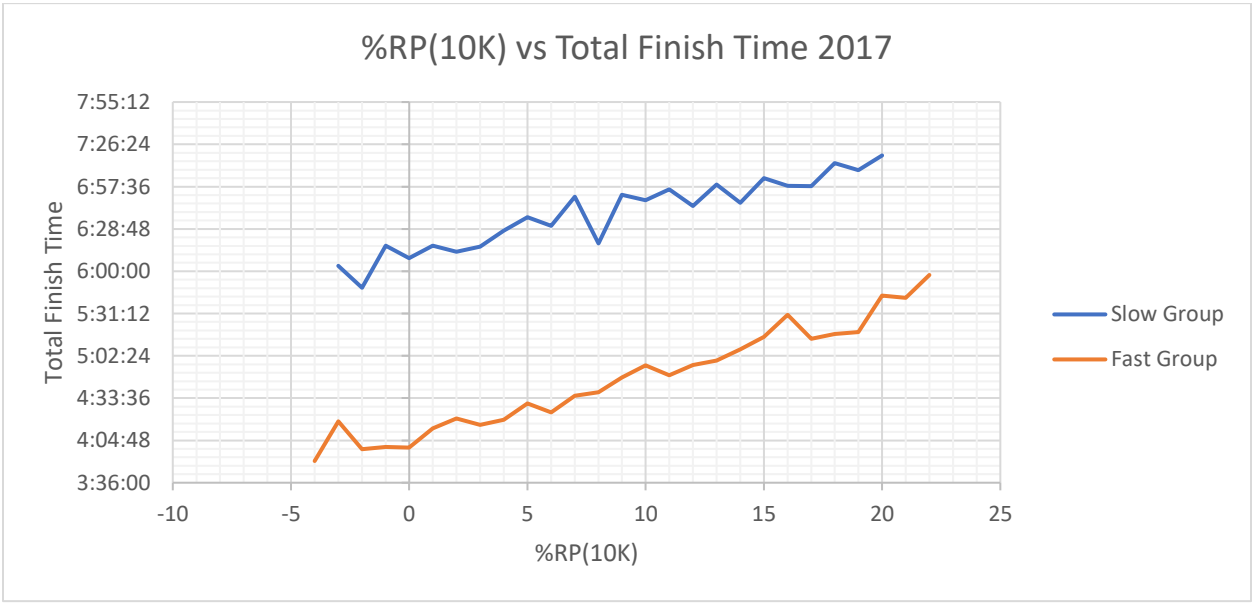

%RP(10K) vs Total Finish Time 2017
Total Finish Time
7:55:12
7:26:24
6:57:36
6:28:48
6:00:00
5:31:12
5:02:24
4:33:36
4:04:48
3:36:00
-10
-5
0
5
10
15
20
25
%RP(10K)
Slow Group
Fast Group

Figure Captions

Figure 1 - Little Rock Marathon Course Map

Figure 2 - Tree-Based Variable Selection

Figure 3 - %RP(10K) vs Total Finish Time 2016

Figure 4 - %RP(10K) vs Total Finish Time 2017