Running the Numbers: The Effect of Age on Performance in Elite Distance Runners

Ben Curtis

May 10th, 2024

ABSTRACT: Professional athletics is often a short career with small purses. Athletes, particularly internationally, start competing at young ages and compete in several events so as to maximize potential career earnings and accolades. Using data from World Athletics 'all-time' rankings, we employed regression analysis to evaluate the relationship between age and performance outcomes in distance runners. We found the optimal performance ages for men and women to be significant, being 24.76 ± 2.33 and 30.85 ± 1.31, respectively. Further, we observed no relationship between event duration and optimal performance age in men and women. This research provides valuable insights into the dynamics shaping athletic achievement in elite distance running, emphasizing the importance of age-specific considerations in talent development and career management.

KEYWORDS: track and field, distance running, optimal performance age, World Athletics, age-performance curve

MANUSCRIPT STATUS: Pre-print

CORRESPONDENCE: curtisbend1@gmail.com

Curtis, B. (2024). Running the Numbers: The Effect of Age on Performance in Elite Distance Runners. In *SportRxiv*. SportRxiv. https://sportrxiv.org/index.php/server/index

I. Introduction

The careers of professional athletes are naturally short. Among elite distance runners, particularly on the international stage, the brevity of professional tenure is accentuated by the rigors of high-level competition and the pursuit of accolades. For all the allure of fame and glory, the financial rewards in distance running are often meager, presenting runners with a narrow window for professional athletics to be financially viable.

Anecdotal evidence suggests a prevailing trend among distance runners to 'move up' in distance as they age, leveraging experience and endurance to tackle longer and more grueling events. Eliud Kipchoge ran a historic marathon world record performance at the age of 38, clocking an astonishing time of 2:01:39. However, the subsequent emergence of young phenoms like Kelvin Kiptum, who at 23 years old eclipsed Kipchoge's mark, underscores the dynamic nature of athletic achievement.

II. Literature Review

The contemporary landscape of research in elite distance running primarily focuses on engaging with the physiological determinants of athletic performance. Scientists have been concerned with evaluating the relationship between age and performance, along with optimal performance ages. For instance, Weippert et al. (2020) elucidated that performance peaks at approximately 24.0 ± 3.0 years for female runners and 23.3 ± 2.6 years for male runners, emphasizing the importance of age as a determinant of athletic excellence. Moreover, empirical literature suggests that quadratic equations best capture the trajectory of individual performances over a career, emphasizing the non-linear nature of age-related effects on athletic achievement (Weippert et al. 2020).

Furthermore, gender-specific differences in the rate of performance decline have been evaluated, particularly in events such as the 800m, 1500m, and 5000m. Ganse et al. (2018) found that the disparity in the rate of performance decline between men and women in these events is not statistically significant, suggesting comparable age effects across genders. This nuanced understanding challenges traditional assumptions about gender-related disparities in athletic performance and underscores the importance of considering event-specific dynamics in age-performance analyses.

Additionally, estimates of peak performance age for distance events provide further insights into the age-related trajectories of elite distance runners. Hollings et al. (2014) reported that the optimal age for peak performance in distance events ranges from 23.9 to 25.5 years for men and 25.2 to 27.4 years for women, highlighting the variability and complexity of age-related effects on athletic achievement. Scholars have advanced our understanding of the complex relationship between age and performance in elite distance running by synthesis of these results, establishing the foundation for future research endeavors that focus on optimal talent development and career management for runners.

III. Data

In order to properly evaluate the relationship between age and performance in elite distance running, we use an 'all-time' performance list for professional track and field. World Athletics, formerly the International Associations of Athletics Federation or IAAF, is the governing body for professionalized track and field athletes and competitions, and curates official performance lists and rankings dating back to the early 1900's. Our dataset is scraped directly from the World Athletics website by Kaggle user 'K04DRUNN3R' and includes observations up until September 17^{th}, 2023. The World Athletics data is particularly useful as age at the time of performance along with several performance metrics such as marks, position results, and results score are included with the original data.

Structurally, our data is pooled data with the unit of observation being a single race performance and the related metrics. After extensive cleaning of the data, removing observations for non-distance running events, field events, and observations where competition age was blank or under 16 years old, there are 107,273 observations. Some observations were excluded for not being pertinent to the question such as with sprinting, field, and relay events, and the others were excluded to prevent regression error by eliminating outliers and missing values. Left for analysis are both conventional Olympic races (such as 800m, 1500m, 5000m), unconventional distances (such as 1000m, Mile, 3000m), and road races from one mile to the marathon.

Tables 1 and 2 present definitions and descriptive statistics for our regressors, offering insights into the characteristics of our dataset comprising 107,273 observations. The mean age of competitors is 25.59 years, with slight variations observed between genders (25.4 years for men and 25.8 years for women). Our analysis is primarily concerned with the results score of a performance. World Athletics created results score for the purpose of ranking and equating performances across events and sexes. For example, a result score of 1243 corresponds to a specific race time or performance mark – an 89.92 meter javelin throw, 19.85 second 200m, or 2:05:07 marathon – for a particular biological sex. As opposed to raw time marks, larger results scores denote better performances. The mean results score is 1154.47, corresponding to a mile time of approximately 3:53.67 for men or 4:25.31 seconds for women. These statistics serve as a foundation for subsequent analyses, informing the interpretation of regression results and the exploration of age-related patterns in elite distance running performance.

IV. Methodology

The following equation is to be estimated via regression analysis:

$$\text{results score} = \beta 1 + \beta 2 \text{ age} + \beta 3 \text{ agesq} + \beta 4 \text{ year} + \text{error}$$

Results score is the outcome variable we are interested in studying, while age and age-squared represent a competitor age and its quadratic, respectively. Inclusion of the quadratic term allows for the exploration of non-linear relationships between age and performance, accommodating potential curvilinear paths of athletic performance within events and careers. Further, we capture the temporal shifts in performance results across competition years with a time-trend variable year. Using a time-trend variable allows our regression to account for overall increases in performance quality as a result of advancements in racing technology or scientific understanding of athletic training.

Our regression employed clustered robust standard errors, addressing two key issues between the data and the assumptions required for ordinary least squares (OLS) regression. First, robust standard errors account for heteroskedasticity that may arise due to the large and diverse nature of the dataset. Second, the recurrence of athletes within the data presents the problem of autocorrelation between performances from the same athlete across several events. Clustering standard errors by an individual numerical code assigned to each competitor, we correct the issue of autocorrelation. Employed in tandem, clustered robust standard errors provide the most efficient and accurate estimates of the regression coefficients.

V. Results

Regression analysis demonstrated that age has a significant effect on the results score of an elite performance. In particular, a significant negative quadratic relationship – illustrated in Figure 1 – between age and performance was observed. Estimates of the regression coefficients and error terms provided in Table 3.1 quantify that negative quadratic relationship and demonstrate significance of the results. Further, when stratified by biological sex, we found the same relationship in both male and female athletes, as per Tables 3.1-a and 3.1-b. Notably, the respective quadratics estimate that age has a greater effect on female athletes and that female athletes peak significantly later than male athletes. Optimal performance age for males was 24.76 ± 2.33 years, and 30.85 ± 1.31 years for females, interval values designating the range in which 95% of elite distance runners will achieve peak performance. Finally, we found no relationship between event distance and optimal performance age in male or female athletes.

Further evaluation of the relationship between optimal age and event duration found that optimal performance age varies between event distance, with some being statistically significant – see Table 3.2. For males, the optimal performance ages for every Olympic event save two were statistically insignificant, the exceptions being the 5000m with an optimal performance age of 26.18 ± 2.62 years and the 800m with an optimal performance age of 29.25 ± 25.39 years. Insignificant results may be caused by linear relationships between age and performance within events, or the much smaller samples used for an event and gender specific regression. Conversely, all calculated optimal performance ages for female athletes were statistically significant. The 3000 meter steeplechase had an optimal performance age of 28.86 ± 1.42 years, an impressively small range given our sample of 976 athletes.

Estimates of average results score by age – as illustrated in Figure 1 – are beneficial for qualitatively observing the relationship between age and performance. However, the variations between best and worst average performances are a mere twenty results score points for both males and females. For a 5000 meter race, an increase of twenty results score points equates to about a five seconds decrease in finishing time: for a 1500m race, a decrease of one and a half seconds. Even though we found age-performance relationship to be statistically significant, the time between best average performance and worst average performances is practically quite small. Variations between best and worst performances were expected to be much less than a smaller longitudinal analysis, as we considered the average elite performance at a particular age over thousands of athletes.

VI. Conclusions

In summary, our analysis successfully quantified and evaluated the relationship between age and performance in elite distance runners. From regression analysis, we found that age had a significant effect on performance outcomes, as measured by the World Athletics results score. A negative-quadratic relationship between age and performance was observed, confirming the existence of optimal performance ages. Furthermore, the preceding results were consistent when stratifying our regression by biological sex. Using our regression coefficient estimates, we constructed age-performance curves for average elite male and female athletic performances. Next, we took advantage of the negative-quadratic relationship to calculate optimal performance ages for males and females in specific Olympic events and overall. Our estimate for optimal male performance age was quite comparable to the established literature – particularly Hollings et al. (2014) and Weippert et al. (2020). Finally, we observed no relationship between optimal performance age and race distance.

Moving forward, the implications of our findings extend beyond the realm of elite distance running, offering valuable insights for athletes and coaches. Understanding the nuanced relationship between age and performance can inform training plans, talent development strategy, and career management approaches in the world of athletics. For athletes, understanding optimal performance ages provides a roadmap for maximizing athletic potential and longevity in the sport. Coaches and trainers can tailor training programs to accommodate age-related performance changes, optimizing performance outcomes.

Furthermore, our study underscores the dynamic nature of athletic achievement and the evolving landscape of elite sports. The emergence of young phenoms like Kelvin Kiptum, who surpassed the marathon world record previously set by Eliud Kipchoge, highlights the continuous evolution and innovation within the sport. As the boundaries of human performance are pushed ever further, our understanding of age-related effects on athletic achievement must remain dynamic and adaptive. By embracing a holistic approach to athlete development and performance optimization, we can ensure that elite distance running remains a vibrant and competitive domain, inspiring generations of athletes to reach new heights of excellence.

Table 1: Variable Definitions

Variable	Definition	Formula or Reference
results score	Official performance ratings calculated and published by World Athletics.	[1]
age	Competitor age	
agesq	Competitor age squared	age * age
date	Date of competition	days after Jan 1st, 1960
gender	Competitor biological sex	Male = 0 Female = 1
runnercode	Unique numerical code assigned to each competitor	

Table 2: Descriptive Statistics

Variable	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
results score	107,273	1154.47	34.32	1000	1306
age	107,273	25.59	4.23	16	46.1
agesq	107,273	647.28	220.02	256	2125.21
date	107,273	17239.22	4933.12	-1612	23720
gender	107,273	0.465	0.499	0	1
runnercode	107,273	5113.35	2990.20	1	10392

Table 3.1: The Effect of Age on Results Score

	(1)	**(2)**	**(3)**
Age	0.36 (0.07)	4.09 (0.70)	4.09 (0.55)
Age Squared		-0.07 (0.01)	-0.07 (0.01)
Date			0.0000008 (0.00007)
R-Squared	0.0021	0.0051	0.0052
Number of Observations	107,273	107,273	107,273

Table 3.1-a: Male

	(1)	(2)	(3)
Age	-0.16 (0.09)	1.40 (0.70)	2.18 (0.70)
Age Squared		-0.03 (0.013)	-0.04 (0.013)
Date			0.00006 (0.00001)
R-Squared	0.0004	0.0008	0.0067
Number of Observations	107,273	107,273	107,273

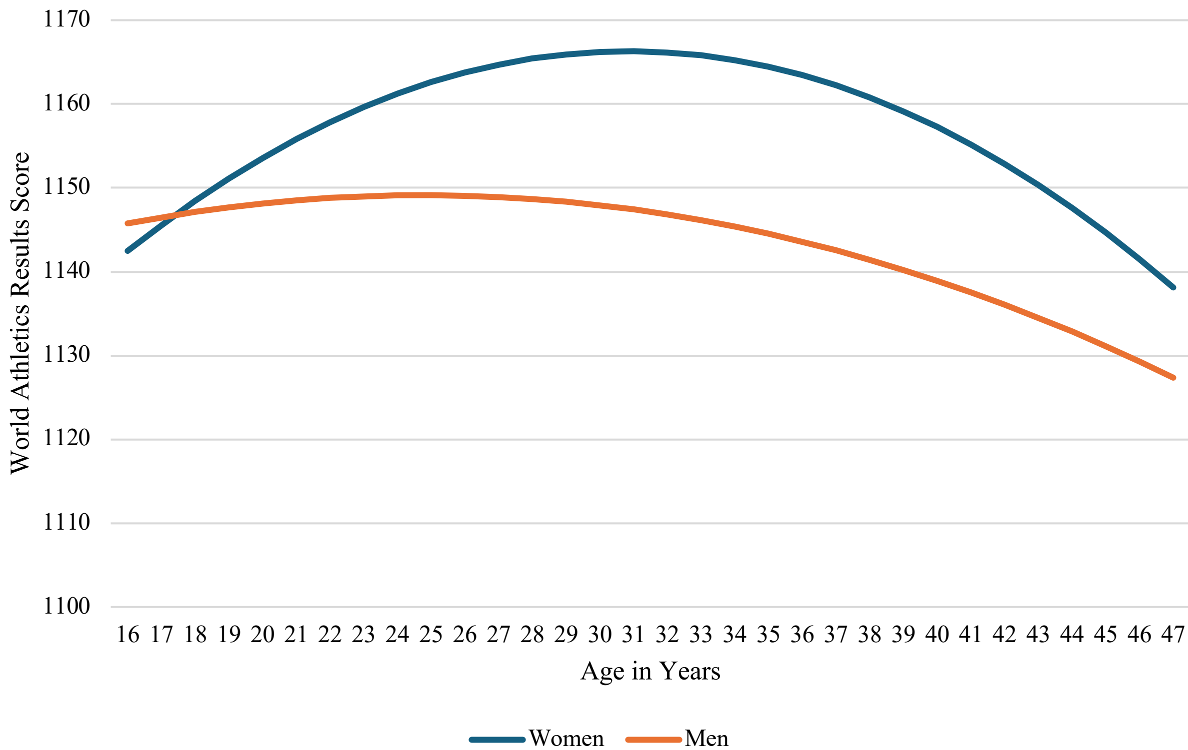
Table 3.1-b: Female

	(1)	(2)	(3)
Age	0.96 (0.104)	6.82 (0.81)	6.67 (0.80)
Age Squared		-0.11 (0.015)	-0.11 (0.015)
Date			-0.0006 (0.0001)
R-Squared	0.0178	0.0268	0.0331
Number of Observations	107,273	107,273	107,273

Table 3.2: Optimal Performance Age by Olympic Event

Event	Male	Female
800 meters	29.25 ± 25.39	30.90 ± 5.19
1500 meters	21.34 ± 32.72	32.44 ± 11.18
3000 meters Steeplechase	43.26 ± 100.5	28.86 ± 1.42
5000 meters	26.18 ± 2.62	25.25 ± 2.14
10000 meters	50.65 ± 105.26	28.04 ± 4.93
Marathon	875.79 ± 77619.06	24.13 ± 5.64
All	24.76 ± 2.33	30.85 ± 1.31

Figure 1: Average World Athletics Results Score by Age



VII. Data Appendix

The regression analysis was conducted using Stata statistical software (version 18.0). The main regression equation estimating the effect of age on performance outcomes was estimated using the "regress" command in Stata. Subsequent regression models stratified by gender were estimated using the same command, with additional controls for potential confounding variables such as year of competition. Clustered standard errors were employed to remove autocorrelation between runners who run multiple events.

Several of the variables required transformation, as the performance marks, dates of birth, and dates of competition were originally stored in 'string' format. These variables were converted to Stata Internal Format, or SIF, and then appropriately formatted the output. SIF stores times as a float representing milliseconds after January 1st, 1960, and dates as an integer representing days after January 1st, 1960. Observations for the half marathon were excluded because of issues converting the string marks to SIF. Due to the large and somewhat redundant nature of the dataset, variables that redundantly denoted event name or birth year and irrelevant variables such as venue and wind measurement were removed.

Data and Stata code used to compute the estimates are available from the author on request.

VIII. References

Ganse, B., Ganse, U., Dahl, J., & Degens, H. (2018). Linear Decrease in Athletic Performance During the Human Life Span. *Frontiers in Physiology*, *9*. https://doi.org/10.3389/fphys.2018.01100

Hollings, S. C., Hopkins, W. G., & Hume, P. A. (2014). Age at Peak Performance of Successful Track & Field Athletes. *International Journal of Sports Science & Coaching*, *9*(4), 651–661. https://doi.org/10.1260/1747-9541.9.4.651

k04dRunn3r. "World Athletics - All-Time Rankings." *Www.kaggle.com*, 17 Sept. 2023, www.kaggle.com/datasets/jeannicolasduval/world-athletics-all-time-rankings. Accessed 9 May 2024.

Weippert, M., Petelczyc, M., Thürkow, C., Behrens, M., & Bruhn, S. (2020). Individual performance progression of German elite female and male middle-distance runners. *European Journal of Sport Science*, 1–7 https://doi.org/10.1080/17461391.2020.1736182