Changepoint Analysis on Marathon and Half Marathon Data
An Insight on Annual Record Trends and Recent Major Marathon Performances

By
Abigail Mabe

Senior Honors Thesis
Statistics and Operations Research
University of North Carolina at Chapel Hill

December 03, 2024

Approved:

Dr. Richard L. Smith, Thesis Advisor

(Dr. William Lassiter), Reader

Abstract

Changepoint Analysis on Marathon and Half Marathon Data
An Insight on Annual Record Trends and Recent Major Marathon Performances

By: Abigail Mabe
(Research supervised by Dr. Richard L. Smith)

Super shoes have been a recent phenomenon in the running community, with many top racers relying on them to boost their running economy. Many biological studies tackle this sensation, suggesting super shoes improve running efficiency. However, is there statistical evidence that these shoes are making a difference in recent marathon performances? In addition, can this recent phenomenon be generalized to observe significant changes in the trend of marathon records over the course of the last few decades? Through the use of Extreme Value Theory, Bayesian statistics, and multiple changepoint models, these questions are tackled in a two-part study. The first part analyzes the historical trends of marathon records, detecting no single event that has changed the trajectory of the fastest marathon times in a given year. The second part analyzes the recent impact of super shoes on marathon performances, establishing a correlation between the introduction of super shoes in 2017 and a significant difference in performance trends.

Acknowledgements

The completion of my honors thesis would not have been possible without the support and guidance of Dr. Richard Smith. With his encouragement and expertise, pursuing research has been one of the best experiences of my undergraduate career. Furthermore, Dr. Richard Smith's dedication and enjoyment of Statistics has inspired my desire to pursue higher education in the subject, for which I am incredibly grateful.

I would also like to extend my sincere thanks to my friends and family for encouraging me throughout the past year. Thank you for your unwavering support and patience as I often enjoy discussing my research with you all.

Table of Contents

List of Figures

1. Introduction

Runners have been competing in documented marathons and half marathons for over 100 years. Until recently, runners have utilized only their training and skills to compete in these races; that is, until the introduction of carbon-plated shoes and other such technologies. Since the introduction of carbon fiber plate shoes in the running industry, companies lobby their use by promoting improved race times. One of these claims comes from a study conducted through the Journal of Applied Physiology. This study affirms carbon fiber-plated shoes lead to marathon finishing times that are 2-3% faster than without the use of the carbon-plated technology [3]. Another comes from the International Journal of Exercise Science, which also alleges carbon-plated shoes improve running economy by 2-4% [9]. One study even goes as far as to say recent improvements in marathon racing must be attributed to technology rather than biological/physiological aspects [2]. However, claims like these are relatively recent and are based on studies with small sample sizes and limited resources, leading to a level of uncertainty in the true effect of these new technologies. Therefore, another method is necessary to explore the possible effects of new technologies on marathon and half marathon racing, which is the motivation for the following analysis.

Part One

Considering general impacts to annual records

Although this analysis in marathon and half marathon records originated from curiosities surrounding the impact of new technologies such as carbon fiber-plated shoes, it can expand to the possibility of many different considerations in annual records. These could include course conditions, locations, or additional external factors, all of which are driving forces in recognizing patterns in annual records. The following analysis focuses on changepoints, or years in which the trend of record-setting significantly changes. This ultimately aims at pinpointing which factor(s) most impact changes in annual record trends, and can help in predicting where records are heading in the future.

2. Data

The data used for the following analysis was manually web scraped off World Athletics. There are four datasets total, one for the women's half marathon, one for the men's half marathon, one for the women's full marathon, and one for the men's full marathon. Each dataset is taken from the *Toplists/All time Top lists* provided by World Athletics [21]. The age category chosen is senior (rather than U18 or U20), provided the assumption most record-holders are expected to be in this age group [1]. The dates which World Athletics hosts records (as of January 2024) spans from 12/31/1899-12/31/2023, although for the marathon and half

marathon races, data is only available spanning back to the 1980s, allowing for approximately 40 years of data for each of the above datasets. Additional options included by World Athletics are *Areas/Countries*, *Best by Athlete/All* and *Limit by Country* which are chosen to be "All" as to not constrict the data in any form. Once these options were selected, manual web scraping was implemented to gather the values residing in the four datasets.

The datasets are all comprised of the singular fastest time from every year, dating back to approximately 40 years ago. To be specific, the women's full marathon dataset dates to 1981, the men's full marathon to 1972, the women's half marathon to 1982, and the men's half marathon to 1987. However, World Athletics' *Toplist* places constrictions on the data available, setting time standards that determine what times can appear on the record. This means any race times above the set limit (determined separately for each race and gender) will not be acknowledged on the *Toplist*. The limit on the men's full marathon is 2 hours, 12 minutes; the limit on the women's full marathon is 2 hours, 30 minutes; the limit on the women's half marathon is 1 hour, 10 minutes. There was no issue with data collection considering these limits, as years were not excluded due to the set standards. However, for the men's half marathon, there was no time in 1986 that was faster than 1 hour, 2 minutes, which is the limit. This means that year was not included in the *Toplist*, and thus not included within the data itself.

Another factor that needs to be taken into consideration when compiling records is race location. For example, the Boston Marathon is no longer a legal record-setting course. This is due to a multitude of reasons, some of which include the fact it is a "point-to-point" course (which can provide a competitor wind assistance) and has a net drop of 459 feet [4]. Despite this, two men's full marathon runners from World Athletics' *Toplist* achieved the fastest annual times there. These runners are Bill Rodgers, who achieved the fastest time of 2:09:27 in 1979, and Robert De Castella, who achieved the fastest time of 2:07:51 in 1986. However, it can be noted that World Athletics' restrictions limiting the validity of courses did not come into fruition until 1990 [4]. Both records cited above occurred before 1990, which is why they justifiably can be included in the data. No other courses provoked an obvious hesitation before considering inclusion in the compilation of the data.

The datasets contain the overall rank of each respective runner in World Athletics' *Toplist* of records, their corresponding time (in hours:minutes:seconds), name, date of birth, nationality, race position, race venue, race date, and results score. However, only the race times and years are required for this type of analysis, so the other features can be discarded for now, although they may be useful in further evaluation. In addition, to make the race times consistent, they are converted from their previous format of hours:minutes:seconds into seconds. The preliminary datasets are displayed below, with time (in seconds)

plotted against date (which is converted to the corresponding year the time was recorded).

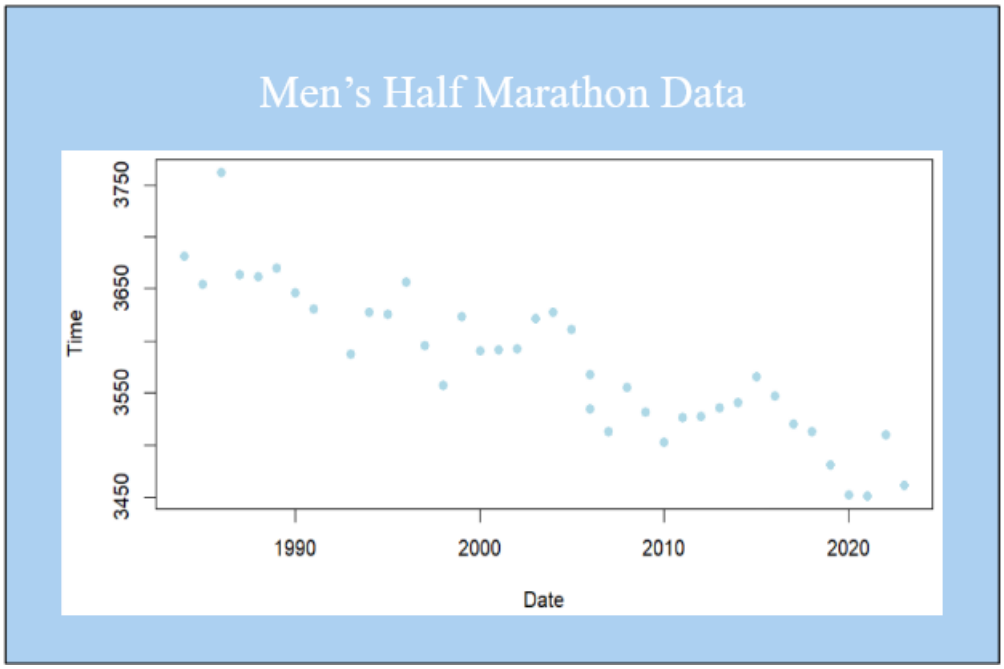


Figure 1: Men's Half Marathon Fastest Times per Year (in seconds)

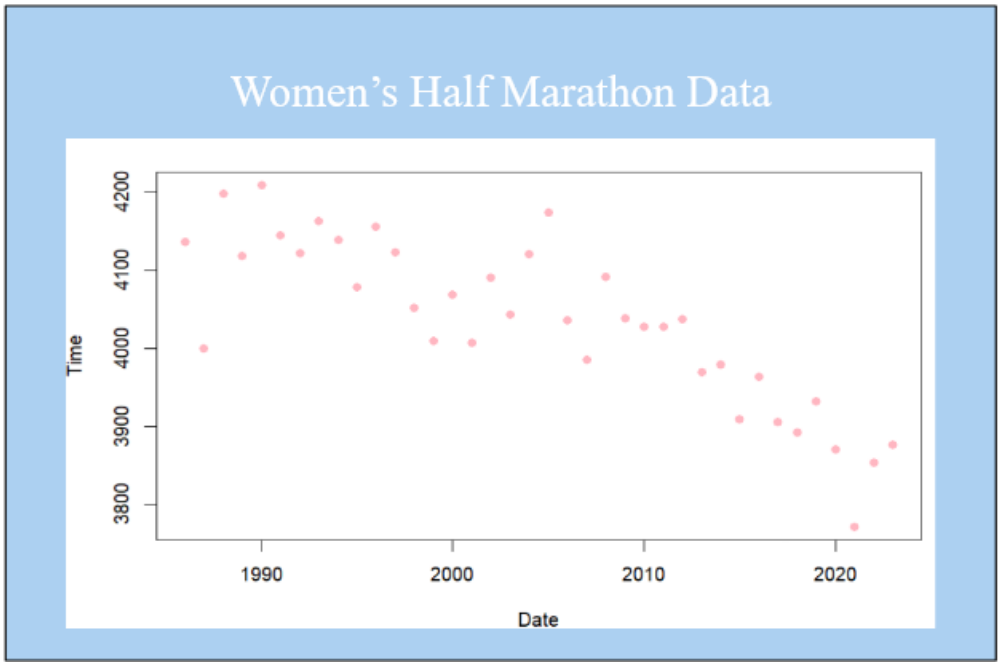


Figure 2: Women's Half Marathon Fastest Times per Year (in seconds)

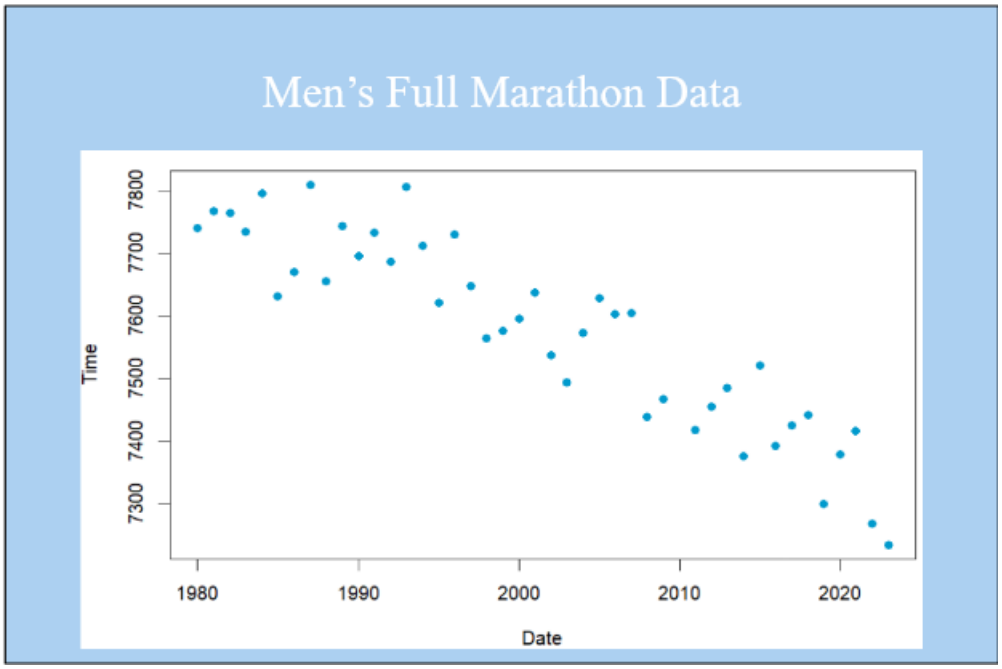


Figure 3: Men's Full Marathon Fastest Times per Year (in seconds)

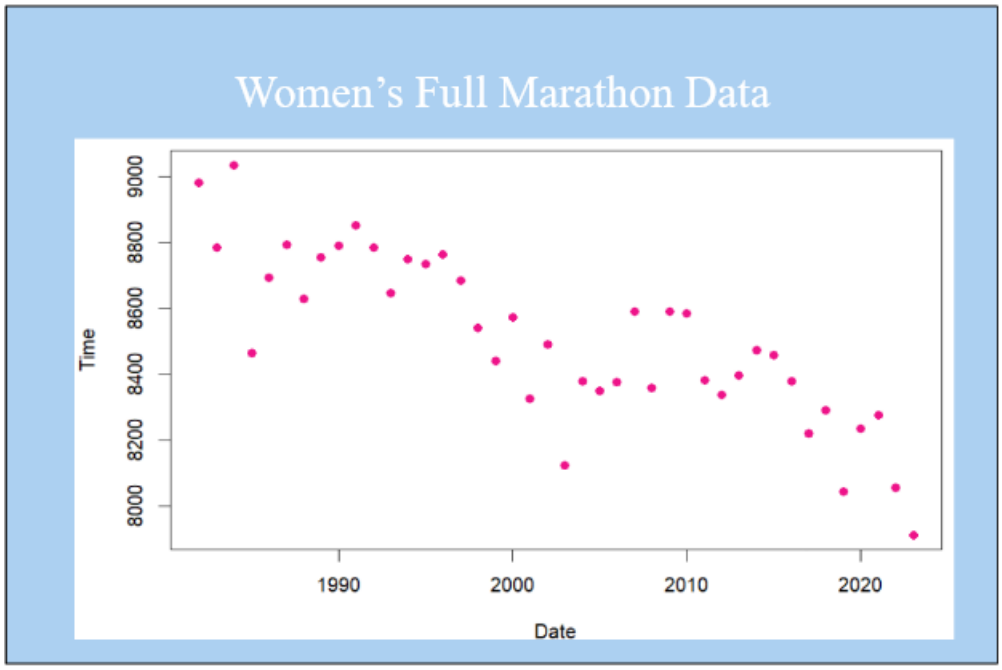


Figure 4: Women's Full Marathon Fastest Times per Year (in seconds)

As observed by the above plots, it can be noticed there is no obvious distribution that can be applied to any particular dataset. Then, estimation methods, discussed in the next section, must be used for further analysis.

3. Methods

First, consider the data from World Athletics, which is a compilation of official marathon/half marathon times on qualifying courses from each year, dating back decades. These times, which can be loosely defined as a sequence of independent random variables $X_1 \dots X_n$, compose M_n, where M_n represents the maximum of the process over n time units of observation. Then, the fastest times from each year are a collection of M_n, although M_n must be reformulated to take the minimum of the recorded times each year, as less time corresponds to a faster race; this concept will be revisited later on.

The goal then is to estimate the distribution of M_n, which must first be done by allowing a linear renormalization, as follows: $M_n^* = \frac{M_n - b_n}{a_n}$, where sequences $\{a_n > 0\}$ and $\{b_n\}$ stabilize the scale and location of M_n^* as n increases. Then, we seek to find the limit distribution for M_n^*, which is given by the extremal types theorem, defined by Stuart Coles in his *An Introduction to Statistical Modeling of Extreme Values* below [8, Theorem 3.1, p. 46]:

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\{\frac{M_n - b_n}{a_n} \leq z\} \to G(z) \text{ as } n \to \infty$$

Where G is a non-degenerate distribution function, G belongs to one of the following families:

$$1) \quad \text{(Gumbel)} \qquad G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \qquad -\infty < z < \infty$$

$$2) \quad \text{(Fréchet)} \qquad G(z) = \begin{cases} 0, & z \leq b, \\ \exp\left[-\left(\frac{z-b}{a}\right)^{-\alpha}\right] & z > b \end{cases}$$

$$3) \quad \text{(Weibull)} \qquad G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{b-z}{a}\right)^{\alpha}\right]\right\}, & z < b, \\ 1, & z \geq b \end{cases}$$

The above three classes of families compose what are known as the extreme value distributions. It can be noted that b and a, which are location and scale parameters respectively, closely reflect sequences a_n and b_n in the linear renormalization of M_n. Additionally, the shape parameter α, applies to Fréchet and Weibull classes of distributions above. Then, given the fact the three types of extreme value distributions above are the only possible limits for the distributions of M_n^*, we obtain an extreme value analog of the central limit theorem.

Although M_n^* must have a limiting distribution that corresponds to one of the three types of extreme value distributions above, each significantly differs from the other. The differences include limit distributions corresponding to upper end points, and behavior of density decay for the three families. The limit distribution G at the upper end point is finite for the Weibull distribution, but approaches ∞ for the Gumbel and Fréchet distributions. In addition, the density of G decays exponentially for the Gumbel distribution and polynomially for the Fréchet distribution. These differences greatly impact the representations of extreme value behavior, so there are obvious weaknesses that correspond to having to choose a family. Another downside to note when choosing a family is a technique is required to choose which family is appropriate, and once a decision is made, an assumption must be followed that the decision is subsequently correct. That is why the generalized extreme value distribution, which can be described as a reformulation of the Weibull, Fréchet, and Gumbel models above, is the best analysis to consider.

Then, the general extreme value theorem can be defined as the following [8, Theorem 3.1.1, p. 48]:

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\{\frac{M_n - b_n}{a_n} \leq z\} \to G(z) \text{ as } n \to \infty$$

For a non-degenerate distribution function G, then G is a member of the GEV family

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\},$$

defined on $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$.

This allows us to model our collection of M_n, or the fastest annual times for marathon and half marathon races. The model has three parameters: a location parameter μ, a scale parameter σ, and a shape parameter ξ. The behavior of these parameters displays the similarity between the GEV and three families of extreme value. When $\xi > 0$, the GEV reflects the Fréchet family and when $\xi < 0$, the GEV reflects the Weibull family. In the case where $\xi = 0$, the GEV reflects the Gumbel family. These observations enforce the idea that the GEV is formulated to include subsets of all three classes of extreme value distribution, solidifying why it is the best choice in analysis.

To re-visit the application of M_n to the data collected, M_n must be reformulated to be the minimum of X_i independent variables, denoted $\tilde{M}_n$. The GEV distribution for minima is similar to the one for maxima, although it has slight differences [8, Theorem 3.3, p. 53]: If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\{\frac{\tilde{M}_n - b_n}{a_n} \leq z\} \to \tilde{G}(z) \text{ as } n \to \infty$$

For a non-degenerate distribution function G, then G is a member of the GEV family:

$$\tilde{G}(z) = 1 - \exp\left\{-\left[1 + \xi\left(\frac{z - \tilde{\mu}}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$$

The above GEV distribution for minima is defined on $\{z : 1 + \xi(z - \tilde{\mu})/\sigma > 0\}$ where $-\infty < \mu < \infty$, $\tilde{\sigma} > 0$ and $-\infty < \xi < \infty$. The GEV distribution for minima can be directly applied to model block minima or indirectly applied by fitting the GEV distribution for maxima to negated data that are realizations from the GEV distribution for minima. In this analysis, the latter was utilized to fit the model, as annual race times (recorded as *Time*) were negated to consistently fit the GEV distribution for maxima.

To verify a GEV distribution will fit the data collected, diagnostic plots can be considered. The first is called a probability plot, which is examined as a comparison of the empirical and fitted distribution functions. More specifically, the probability plot considers the empirical distribution function evaluated at $z_{(i)}$ when data $z_{(1)} \leq \ldots \leq z_{(m)}$ are ordered block maxima. This expression is given by: $\tilde{G}(z_{(i)}) = \frac{(i-0.5)}{m}$, where the corresponding fitted distribution function is specified as follows:

$$\hat{G}(z_{(i)}) = \exp\left\{-\left[1 + \hat{\xi}\left(\frac{z_{(i)} - \hat{\mu}}{\hat{\sigma}}\right)\right]^{\frac{-1}{\hat{\xi}}}\right\}$$

In the case where the GEV model works well, it is expected that $\hat{G}(z_{(i)}) \approx \tilde{G}(z_{(i)})$ for each i. Then, a probability plot consisting of the points

$$\left\{(\tilde{G}(z_{(i)}), \hat{G}(z_{(i)}), i = 1, \ldots, m\right\}$$

should display a visualization that is similar to a unit diagonal; otherwise, there could be failure in utilizing the GEV model for the specified data. Although this plot establishes a way to assess the fit of the GEV model, additional diagnostic plots should be considered, especially given the fact that a probability plot provides the least information in the region of most interest; as $z_{(i)}$ increases, $\hat{G}(z_{(i)})$ and $\tilde{G}(z_{(i)})$ are bound to approach 1. This issue is avoided by the quantile plot, another diagnostic plot, utilized to ensure the fit of a specified model to data.

In the instance where the specified model is a GEV model, the quantile plot consists of the points

$\left\{(\hat{G}^{-1}((i-0.5)/m), z_{(i)}),\ i = 1, \ldots, m\right\}$, where

$$\hat{G}^{-1}\left(\frac{i-0.5}{m}\right) = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - \left\{-\log\left(\frac{i-0.5}{m}\right)\right\}^{-\hat{\xi}}\right].$$

Similar to the above probability plot, substantial departures from linearity in the quantile plot indicate failure in the GEV model for the specified data. It can be noticed that the quantile plot mirrors the probability plot closely in statistical origination, which is why it was important to specify the probability plot first, although the probability plot diagnostic is not as common as the quantile plot diagnostic. In addition to the probability and quantile plots, a particularly convenient diagnostic for interpreting extreme value models is the return level plot.

The return level plot consists of the points $\{(log\, y_p, \hat{z}_p) : 0 < p < 1\}$, where $\hat{z}_p$ is the maximum likelihood estimate of z_p, plotted on $z_p = \mu - \frac{\sigma}{\xi}[1 - \{-log(1-p)\}^{-\xi}]$ and $y_p = -log(1-p)$. This results in a plot that displays return level estimates for long return periods, again allowing the linearity of the plot (when $\xi = 0$) to provide a baseline of the goodness of fit of the GEV model on the data. When combined with empirical estimates of the return level function, the model-based curve and empirical estimates should be in reasonable agreement; otherwise, caution should be taken when proceeding with the GEV model. An equivalent, but generally less informative, diagnostic displays the probability density function of a fitted model with a histogram of the data. Nevertheless, it is important to consider multiple diagnostic plots to ensure a holistic understanding of how well the desired model fits the data.

Viewed below are the diagnostic plots of the gathered data for men's half marathon. The top left is the quantile plot described above, which plots empirical quantiles against model quantiles. A good fit will yield a straight line of points, which can be observed in the visualization. The top right is a slightly modified quantile plot, with different values of simulated data from the fitted distribution function, plotted along with a one-to-one and regression line fitting the quantiles. This data also seems to be consistent with a straight line of points, indicating the data might be a good fit for the GEV model. The bottom left plot displays the less informative histogram diagnostic, where the blue dashed line displays the model density, and the solid line displays the general histogram of the data; although they seem to be comparable, this is still a less informative graphic than the others due to the grouping style of histograms. The bottom right plot displays the return level plot described above, along with plotted 95% confidence intervals shown as gray dashed lines. The position of the plotted points on the y-axis correspond to values of observed maxima in the data, which is comparable to the best estimate, or the solid line on the plot. Then, due to the linear nature of the quantile plots, and the comparable nature of the density and return level plots, it can be concluded the

GEV model is suitable for the men's half marathon data.

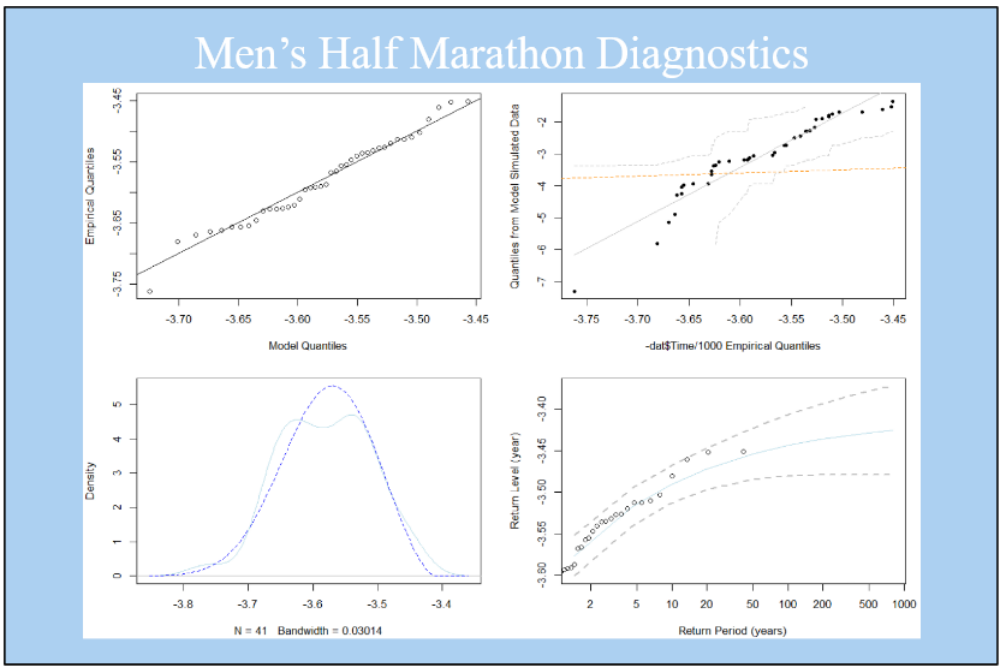


Figure 5: Men's Half Marathon Diagnostic Plots.

The men's full marathon, women's full marathon, and women's half marathon diagnostics all display similar patters to the diagnostic plots for the men's half marathon records, which can conclude the diagnostic portion of ensuring the GEV model fits the gathered data.

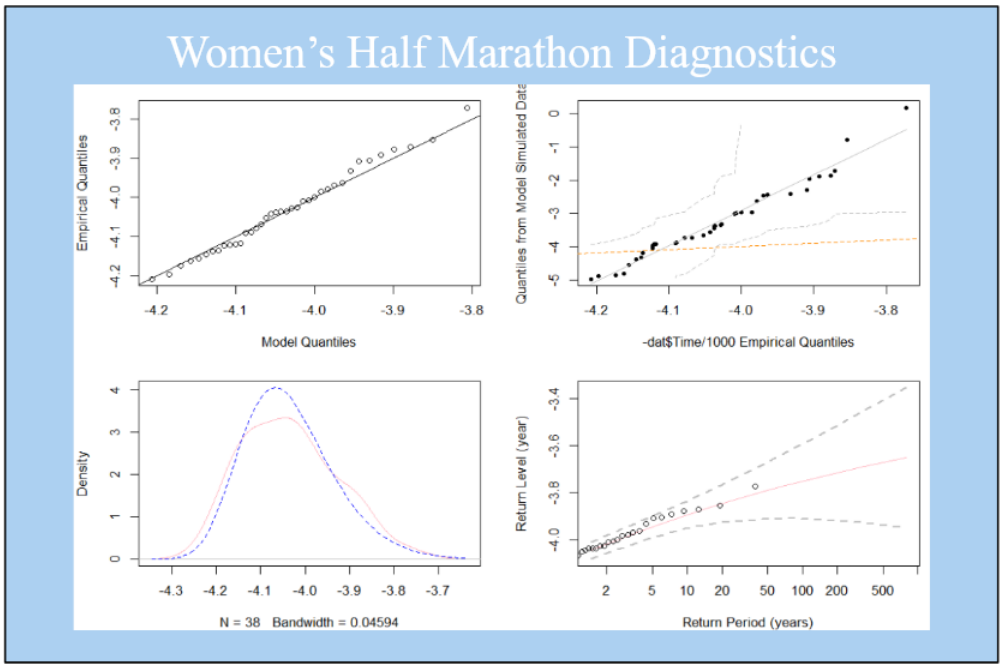


Figure 6: Women's Half Marathon Diagnostic Plots.

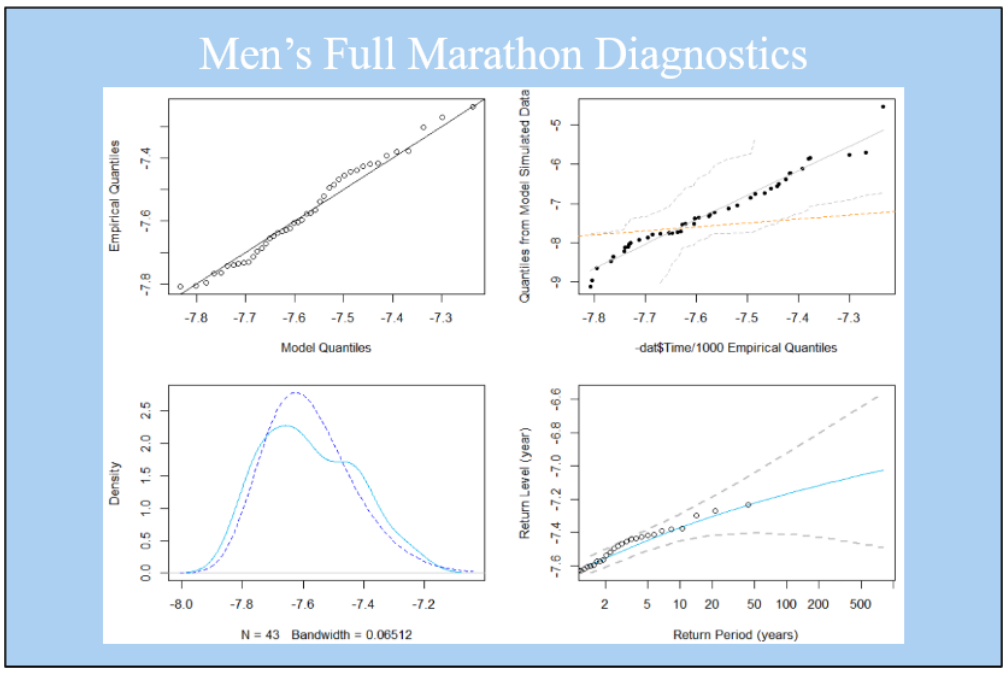


Figure 7: Men's Full Marathon Diagnostic Plots.

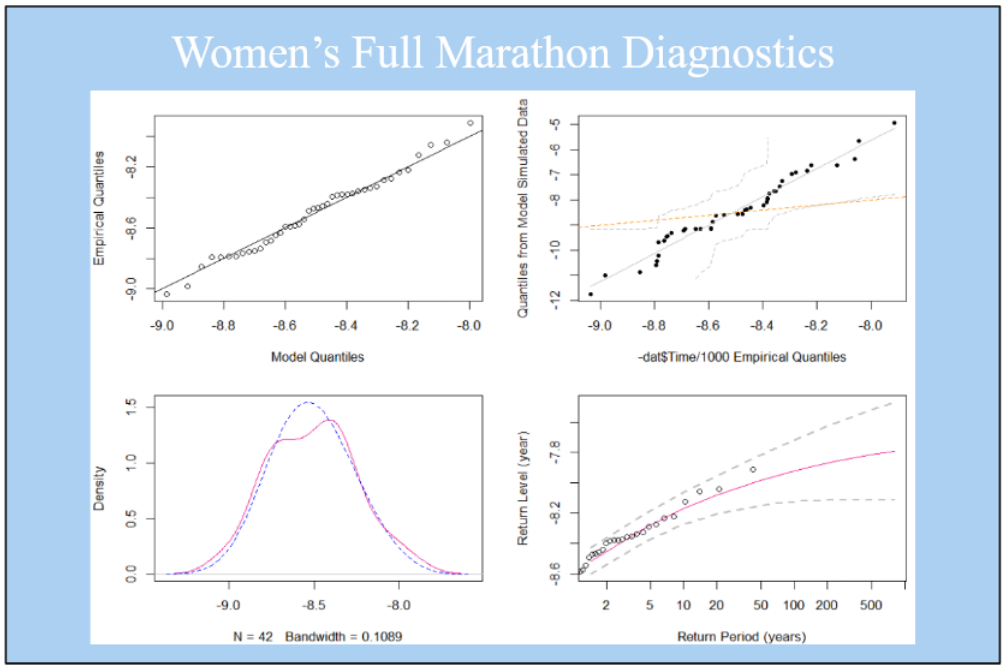


Figure 8: Women's Full Marathon Diagnostic Plots.

Now that the general model has been established for the collected data, the details of the analysis can be considered. To examine whether events have caused significant changes to the trend of annual marathon/half marathon records, a changepoint analysis was considered appropriate. To begin, there must

be an assumption that the data is a non-stationary process, or a process that has characteristics that change systematically through time. Then, utilizing the GEV distribution for maxima outlined above (with parameters μ, σ, and ξ), a suitable model can be considered for Z_t; the fastest times from each year in marathon or half marathon races, as follows:

$$Z_t \sim GEV(\mu(t), \sigma, \xi), \text{ where } \mu(t) = \beta_0 + \beta_1(t) + \beta_2(t - t_1)_+$$

and $(t - t_1)$ represents a changepoint in the trend at time t_1. This can be re-modeled in the following way, as pertains to Stuart Coles [8, p. 106]:

$$\mu(t) = \begin{cases} \beta_0 + \beta_1(t) \text{ for } t \leq t_1, \\ \beta_0 + \beta_1(t) + \beta_2(t - t_1)_+ \text{ for } t > t_1 \end{cases}$$

Although both expressions are equivalent, for the simplicity of modeling the first formulation is utilized in the analysis of this data. Other extreme value parameters can be modeled in a similar manner, but since the goal is to find the location of a possible changepoint in the data, μ is the best choice over σ or ξ.

To evaluate the model $GEV(\mu(t), \sigma, \xi)$, a maximum likelihood estimation can be utilized. Generally speaking, maximum likelihood is a method of estimation of a vector of unknown parameters θ within a specified family. The probability of the observed data as a function of θ is the likelihood function, where values of theta that have a high likelihood correspond to models which give high probability to the observed data. The idea behind maximum likelihood is to adopt the model with the greatest likelihood, given that is the model that assigns the highest probability to the observed data. Then, in the case of the GEV distribution, the maximum likelihood estimation is described as follows [8, Equations 3.7-3.8, p. 55]:

Under the assumption that $Z_1, \ldots, Z_m$ are independent variables having the GEV distribution, the log-likelihood for the GEV parameters when $\xi \neq 0$ is

$$L(\mu, \sigma, \xi) = -m \log \sigma - (1 + 1/\xi) \sum_{i=1}^{m} \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m} \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1/\xi},$$

provided that $1 + \xi \left(\frac{z_i - \mu}{\sigma}\right) > 0$, for $i = 1, ..., m$.

If $\xi = 0$, a different log-likelihood estimator is utilized than what was described above. In theory, numeric optimization could converge close to 0, in which case the limiting Gumbel form of the GEV could be substituted into the model, However, in practice, this never happens, so for this analysis, the separate formu-

lation of the log-likelihood estimator (using the Gumbel limit of the GEV distribution) can be disregarded, allowing the log-likelihood estimator outlined above to be utilized for the analysis in its entirety. One detail to note in the above formulation is the fact the analysis is assuming a non-stationary distribution of $\mu(t)$, so current values of μ will subsequently be replaced by $\mu(t)$ to account for this assumption in changepoint modeling.

The application of the log-likelihood estimator previously mentioned comes in the form of a built-in function from the extRemes package in R [10]. This function mimics the same formulation of the log-likelihood estimator considering GEV parameters when $\xi \neq 0$, although the scale parameter is exponentiated in the function's output. One very important issue that concerns the computation of such likelihoods is the aspect of scaling. In the case of the data utilized for this analysis, both *Date* and *Time* had to be re-scaled to gather accurate log-likelihood estimates. Given *Time* is the number of seconds in each annual record, this variable contains numbers that are consistently in the thousands. Additionally, as *Date* is recorded in years such as 2003, this variable also contains numbers that are inadvertently recorded as thousands. High numbers that are in the thousands typically do not perform as well under the maximum log-likelihood estimator, especially through computations by the likelihood function built into an R package. It should be noted there is no theoretical reason the algorithm should not converge, regardless of the scale of the inputs. In practice, however, these algorithms become numerically unstable with large orders of magnitude difference between variables Then, rescaling of the data was necessary, where *Time* was divided by 1000 and *Date* was divided by 100 to account for these considerations.

The output of the built-in likelihood estimator function (named fevd) includes a negative log-likelihood estimate, as well as corresponding parameter estimates. It is important to note that minimizing the negative log-likelihood is equivalent to maximizing the log-likelihood, although in many cases it is easier to perform optimization by minimizing a function rather than maximizing one. This is why computing the negative log-likelihood is more common in statistical packages and operations, and that is the method of optimization utilized in this analysis.

To model the data as a non-stationary GEV process with a changepoint, there needs to be a way to estimate where that changepoint occurs in the location parameter $\mu(t)$. Recall $\mu(t) = \beta_0 + \beta_1(t) + \beta_2(t - t_1)_+$, where t_1 indicates the location of the changepoint. Then, to estimate the true value of t_1, the negative log-likelihood for each possible value of t_1 is computed through fevd and stored for future computation.

In addition, a negative log-likelihood value is computed from the linear trend of $\mu(t)$, given by $\mu(t) = \beta_0 + \beta_1(t)$. The linear and changepoint negative log-likelihood values are compared through the following method, known as the likelihood ratio test [22, Equations 8.18-8.19, p. 132]:

Consider a hypothesis test in the form of a null hypothesis $H_0 : \theta_1 = \theta_1^0, \ldots, \theta_k = \theta_k^0$, where θ forms an open subset of R_d, and the alternative hypothesis H_1 with unrestricted $\theta_1, \ldots, \theta_d$. Also, note $1 \leq k \leq d$ and $\theta_1^0, \ldots, \theta_k^0$ are known values. Then, letting $L(\theta)$ denote the likelihood function and Θ_0 denote the subset of Θ constrained by the null hypothesis above, we have

$$L_0 = sup\{L(\theta) : \theta \; in \; \Theta_0\}, \quad L_1 = sup\{L(\theta) : \theta \; in \; \Theta\},$$

where the resulting likelihood ratio statistic is formulated as: $T_m = 2\, log\left(\frac{L_1}{L_0}\right)$, given m denotes sample size.

In this formulation, L_1 is the negative log-likelihood of the changepoint model for each respective value of t_1 in $\mu(t)$, where L_0 is the negative log-likelihood of the linear trend for $\mu(t)$, which is constant among t_1 values. The reason behind using the likelihood ratio test is to find which value of t_1 in the changepoint model establishes the greatest distance from the "null" model, or the model fitted to a linear trend. Then, the greatest value of the likelihood ratio statistic (or the individual values computed from the likelihood ratio test) in the iterations through all possible values of t_1 likely indicates the true value of t_1 in the data.

As these likelihood ratio statistics were created for each possible t_1 in the data, the number of observations should be as long as the data itself. To visualize these values better, the likelihood ratio statistics were plotted for each possible value of t_1 in the following graphics. The *Index* of the maximum point indicates the most probable true value of t_1. The smaller values of *Index* correspond to more recent years (with 2023 being the most recent, corresponding to an *Index* of 0), and the larger values of *Index* correspond to earlier years. Below are the plots that display the likelihood ratio statistics for the data collected, with the likelihood ratio statistics displayed on the y-axis and their corresponding *Index* on the x-axis.

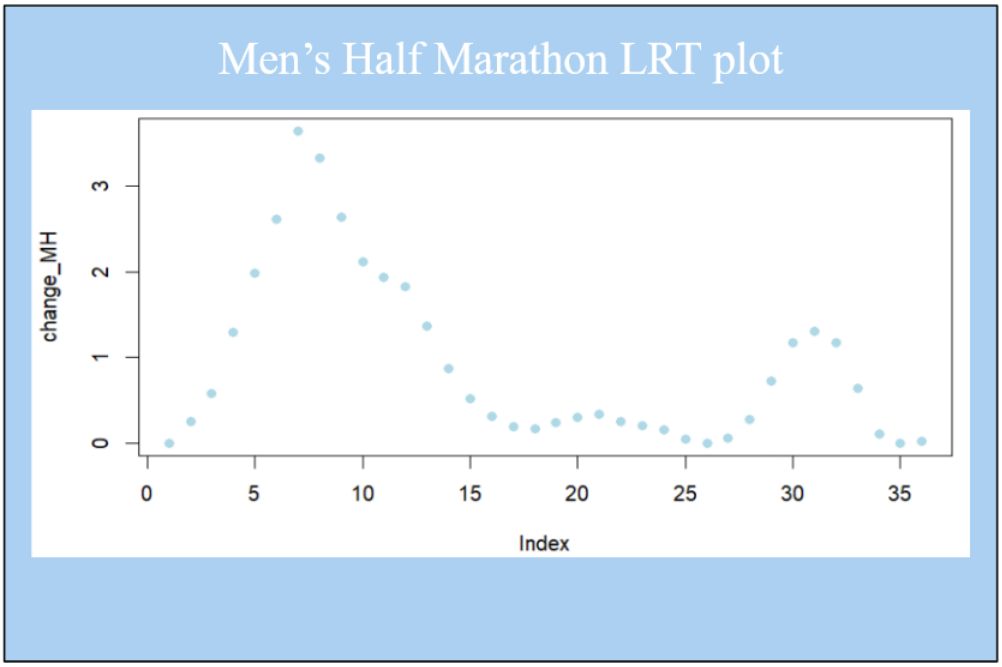


Figure 9: Men's Half Marathon LRT plot

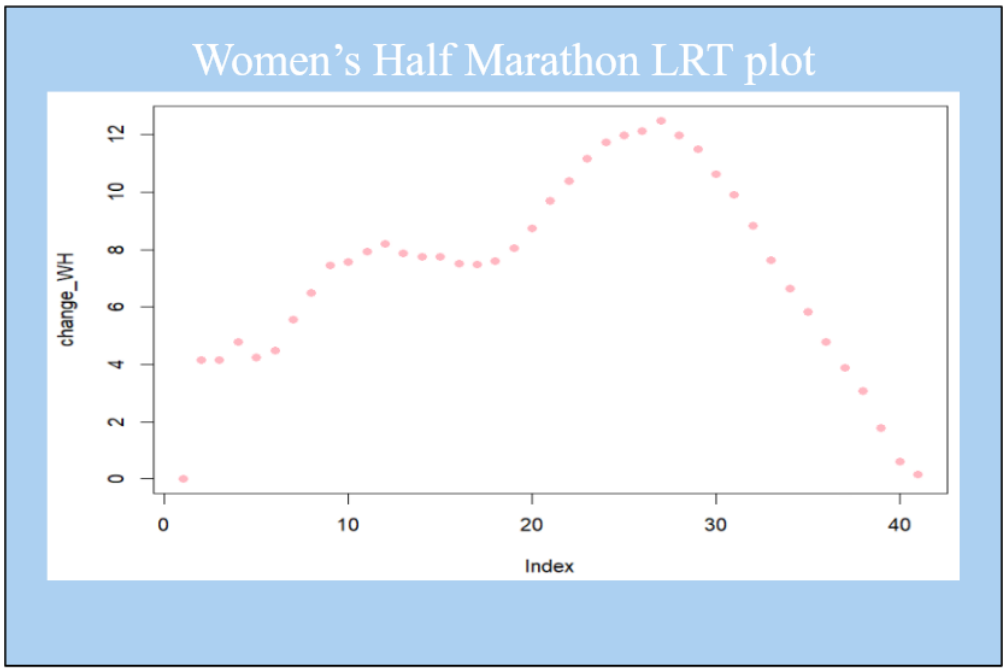


Figure 10: Women's Half Marathon LRT plot

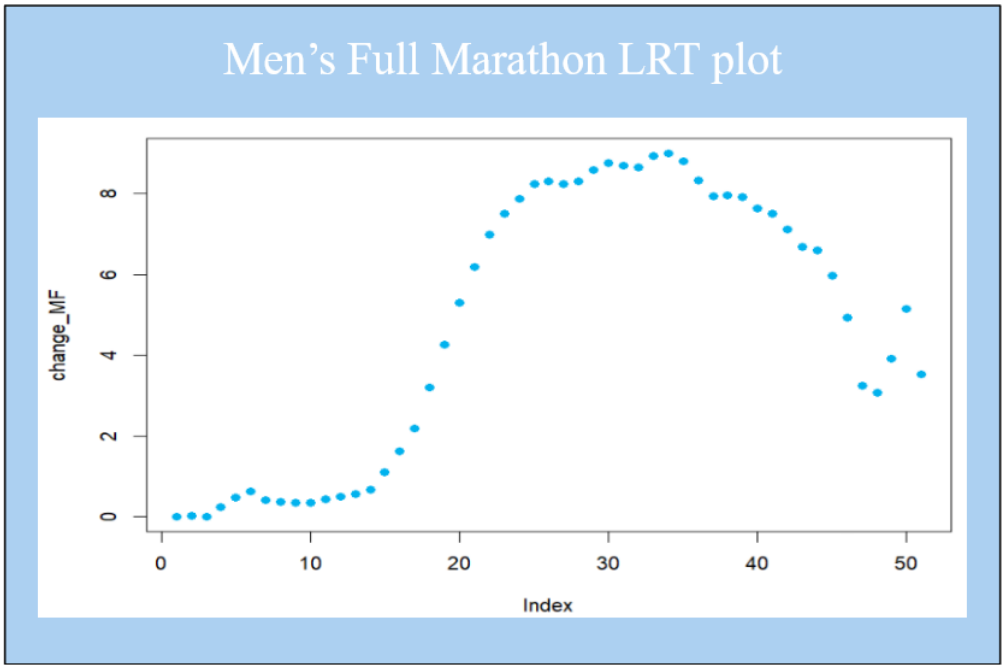


Figure 11: Men's Full Marathon LRT plot

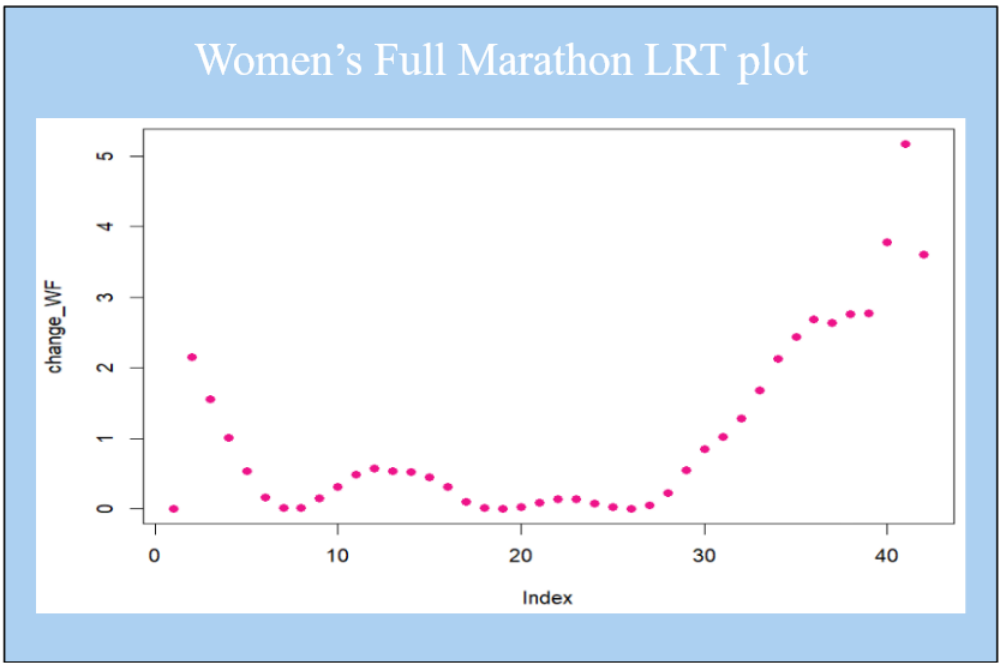


Figure 12: Women's Full Marathon LRT plot

These plots will be referenced later when discussing results, but a few observations should be noted to motivate the rest of the analysis. These include the fact that both the women's half marathon and the men's full marathon have what seem to be clear changepoints at a time t_1 that are present outside of the endpoints

of the given data. In contrast, the women's full marathon shows no such trend and the men's half marathon shows what is likely a two changepoint model, which is not what the current analysis is examining.

To prove the values of the most probable respective changepoints t_1, observed above, a hypothesis test of sorts must be conducted, especially to claim statistical significance of said changepoints. This can be done through a simulation of datapoints paired with an approximate hypothesis test derived from likelihood ratio statistics.

To compute the hypothesis test outlined above, random simulations must be drawn from the linear GEV distribution for each dataset and fit to both the changepoint and linear models of $\mu(t)$. Then, the negative log-likelihood is found for the linear model of $\mu(t)$, as well as for each possible value of t_1 in the changepoint model, both of which are fit to the simulated data. After computing negative log-likelihoods, the largest of the individual likelihood ratio statistics are kept to comprise what should be normally distributed values. Ideally, $1,000$ simulations of this occurs for each dataset, and the maximum likelihood ratio statistic stored for each of the simulations. These values compose the data used to calculate the test statistic, along with the value of the maximum likelihood ratio statistic calculated from real data.

One important point to consider is the distribution of the simulated likelihood ratio statistics. Typically, a likelihood ratio statistic resembles an asymptotic chi-squared (χ^2) distribution, where $T_n \xrightarrow{d} \chi^2_m$ given H_0 is true as $n \to \infty$ (Young, p. 132). However, in the case of change-point inference, these assumptions of distributions are generally false. Rather, the limiting processes calculated from possible likelihood ratio statistics typically have non-differentiable sample paths. This could be from the obvious dependency between likelihood ratio statistics computed across all values of t_1, as in the case of this analysis, or the suggestion that detailed results are confined to stationary Gaussian processes [17, p. 361]. Some ideas of what the actual change-point likelihood ratio statistic distribution resembles are based off methods of sequential analysis, or one derived from a general exponential family [17, p. 362].

In lieu of the unusual nature of likelihood ratio test distributions associated with change-point models, many approximate hypotheses test suggestions have been introduced. These include proposals of conditional and Bayesian approaches, as well as an approach called the Davies approach [17, p. 363]. Nevertheless, since the computed likelihood ratio statistics from the above simulation do not reflect a typical chi-squared distribution, the approximate hypothesis test must be based on the distribution given by each respective simulation. The method utilized in this analysis works in the following fashion: the aim is to see how probable it is for a given simulated likelihood ratio statistic to be greater than the actual likelihood ratio statistic for each respective dataset. Then, the proportion of simulated likelihood ratio statistics that are greater than the actual likelihood ratio statistic must be calculated to estimate significance. If that proportion for a dataset

is less than 0.05, it can be said that dataset has a statistically significant changepoint. However, there should be caution when assigning a statistically significant changepoint for a dataset that has a proportion greater than 0.05, and no statistical significance should be assigned to a changepoint in a dataset with a proportion greater than 0.10. Then, the analysis is concluded for a one-changepoint model. Further discussion will follow in the results.

4. Results

As mentioned above, after observation of the likelihood ratio plots, it can be hypothesized that the men's full marathon and women's half marathon data contain evidence for a changepoint. However, these observations can only predict a likely changepoint, so simulations and formal hypothesis testing must be considered. The results of the simulation, conducted on all datasets, are displayed in the following plots. The plots have multiple components, including a histogram and analogous smoothed line displaying the density of the likelihood ratio statistics computed on the simulated data. In addition, a vertical line is plotted on top of this histogram, representing the likelihood ratio statistic for the actual data:

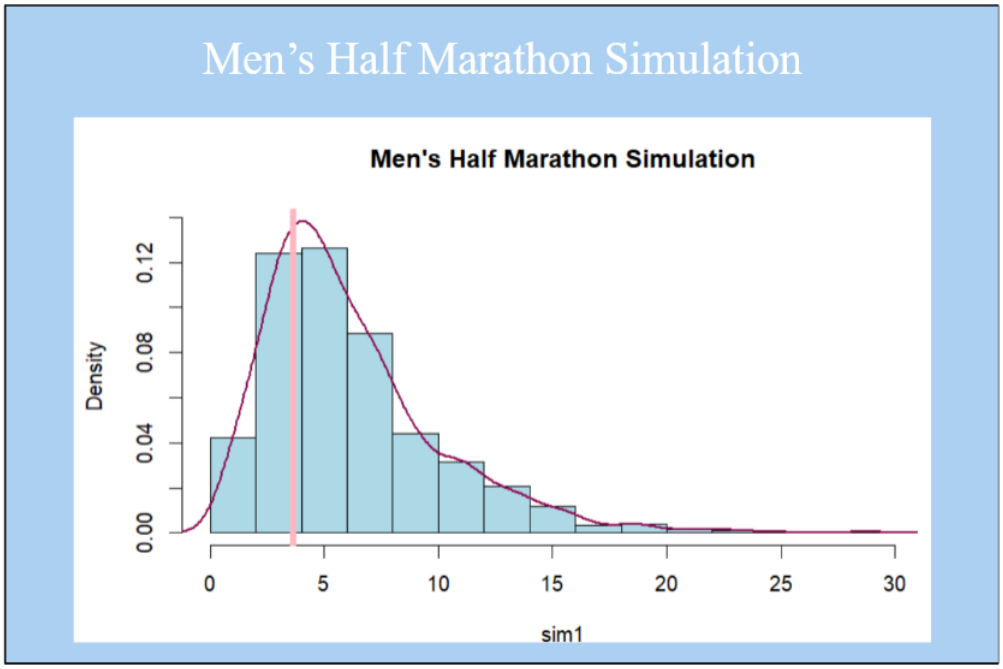


Figure 13: Men's Half Marathon LRT Simulation and Statistical Examination

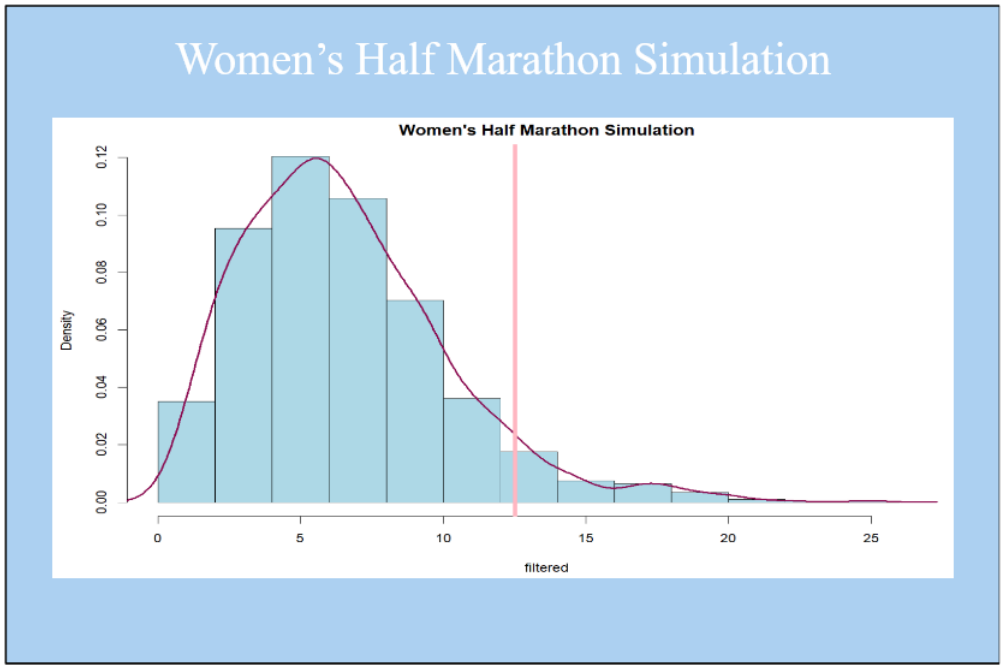


Figure 14: Women's Half Marathon LRT Simulation and Statistical Examination

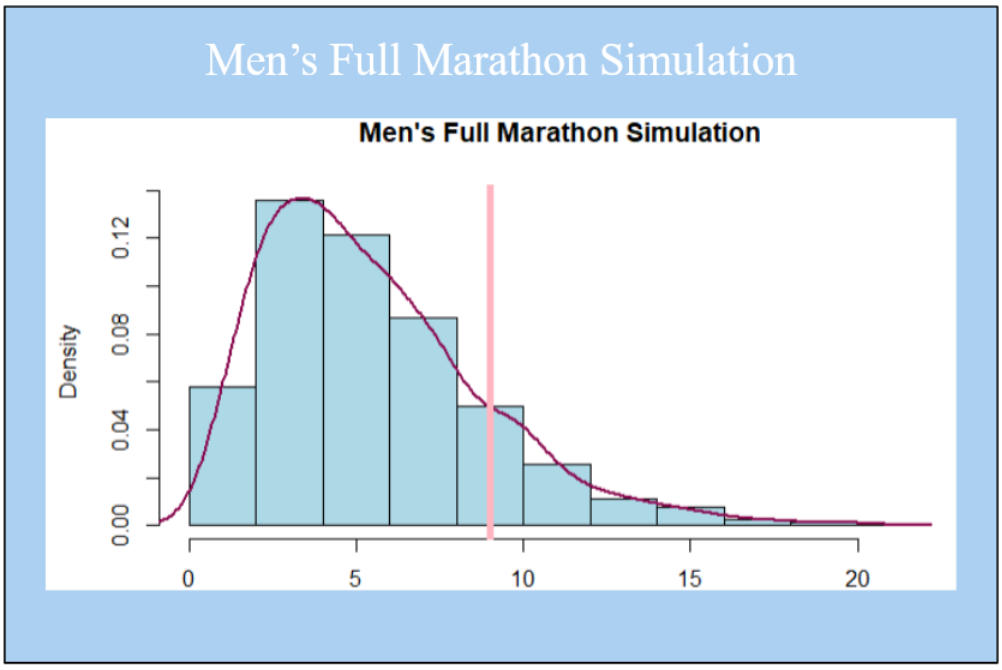


Figure 15: Men's Full Marathon LRT Simulation and Statistical Examination

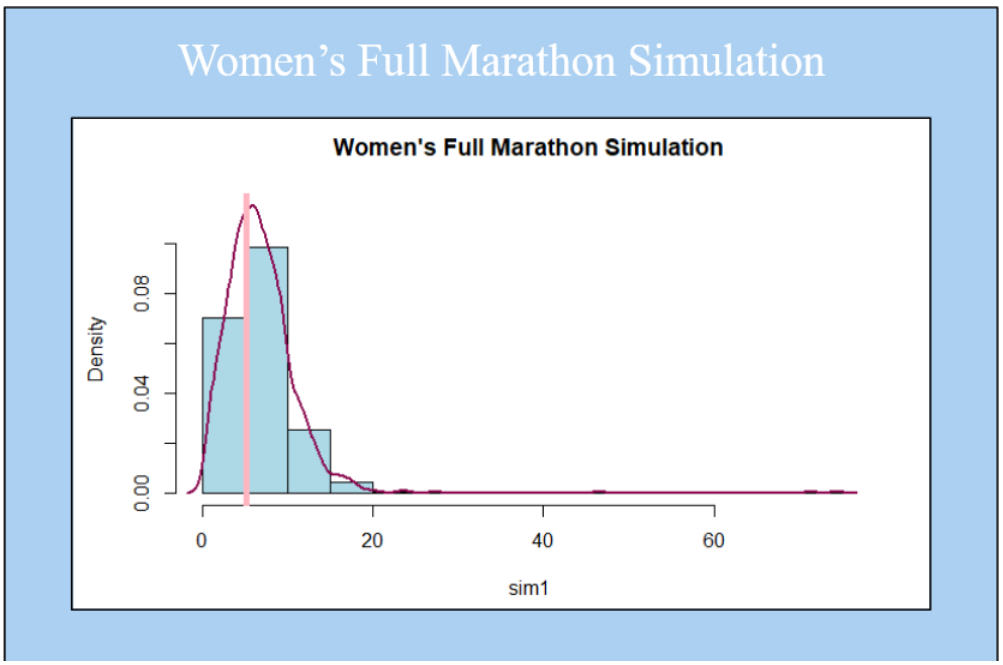


Figure 16: Women's Full Marathon LRT Simulation and Statistical Examination

These plots serve as further proof that there is a likely changepoint in the women's half marathon data, although it seems less likely for there to be a significant changepoint in the men's full marathon data. This is given by the fact that the vertical line representing the likelihood ratio statistic for the women's half marathon seems to be greater than approximately 95 of the simulated likelihood ratio statistics displayed above. However, the vertical line representing the likelihood ratio statistic for the men's full marathon seems to only be greater than approximately 85 of the simulated likelihood ratio statistics displayed above. To formally confirm this observation, an approximate hypothesis test was conducted on the simulated data displayed in the plots above. The results of each test are shown below, where the proportion of simulated likelihood ratio statistics greater than the actual likelihood ratio statistic is calculated for each test:

Men's Half Marathon : 0.772

Women's Half Marathon : 0.065

Men's Full Marathon : 0.15

Women's Full Marathon : 0.631

It can be noted that the above numbers represent the probability that a simulated likelihood ratio

statistic is greater than the actual likelihood ratio statistic, indicating whether there are statistically significant changepoints in each dataset. Then, there is weak evidence for a changepoint in the women's half marathon data given the test probability is 0.065; this is not quite strong enough to be statistically significant (< 0.05) but is not weak enough to establish there is no evidence for a changepoint whatsoever (> 0.10). However, the other datasets did not establish evidence for a changepoint model, as the probabilities displayed above are all > 0.10.

The changepoint supported by weak evidence for the women's half marathon occurs at the index of which the maximum likelihood ratio statistic is recorded, so in the year 1996. Then, the plot for the women's half marathon, fit to a changepoint model with linear trends on either side of the changepoint, is displayed below:

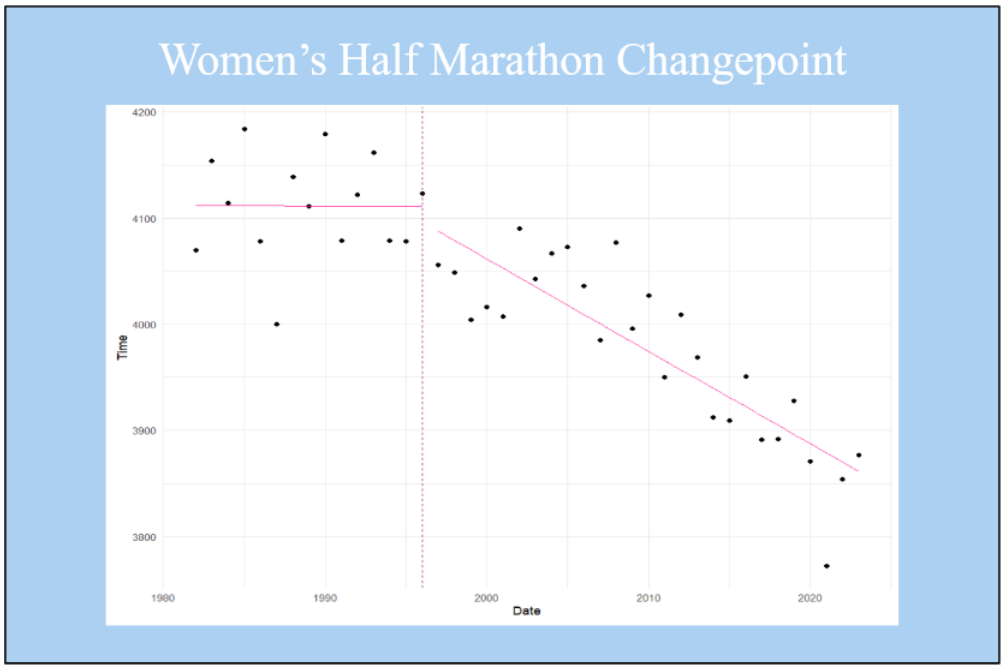


Figure 17: Women's Half Marathon Annual Records fit to a Changepoint Model

5. Discussion

As noted above, the only dataset with any evidence for a changepoint is the women's half marathon, where there is weak evidence for a changepoint in 1996. This means the trend in women's half marathon annual records before and after 1996 are likely different, which could be due to a multitude of reasons. The first consideration in determining these differences in trends should be looking at the racers themselves.

In 1996, the first Kenyan runner in the dataset, Tegla Loroupe, set the annual record for the women's

half marathon in 1 hour, 08 minutes, and 43 seconds (1 : 08 : 43). After Loroupe, Elana Meyer had a 3-year annual record streak where she became the first woman to break 1 hour, 07 minutes (1 : 07 : 00) in the half marathon. The annual records following Meyer's 3-year streak resembled similar patterns and times, plateauing again until Mary Jepkosgei Keitany broke 1 hour, 06 (1 : 06 : 00) minutes in the half marathon in 2010. After Keitany's accomplishment, annual records have become increasingly faster, with the current record of 1 hour, 02 minutes, and 52 seconds (1 : 02 : 52) achieved by Letesenbet Gidey in 2021. However, prior to Loroupe in 1996, times consistently toggled between 1 hour, 08 minutes (1 : 08 : 00) and 1 hour, 09 minutes, and 30 seconds (1 : 09 : 30). This could be explained by the nature of long-distance races, where records accompany a pattern of constant times, followed by a period of faster times. However, there are other aspects to consider besides the prevailing nature of long-distance racing.

Although these other factors are less convincing, they are still worth mentioning for consideration. This could include the fact that half marathon races took place in later years, such as the World Athletics Half Marathon Championships that first took place in 1992 [19]. Although these events were introduced in the early 1990s, they likely took a few years to gain popularity as major events, which could explain the sudden change in the annual records trend starting in 1996. In addition, women's long-distance events became more popular in the late 1970s/early 1980s with the first addition of the women's Olympic marathon in 1984. These two factors likely intersected to create popularity in the women's half marathon around 1996, where there is weak evidence for a changepoint. Then, consideration should be given towards some combination of these factors along with the existing nature of long-distance racing.

6. Conclusion

The changepoint analysis outlined above gives considerable thought towards the "trends" of long-distance running, one of recent being carbon fiber-plated shoes. As the analysis showed, there is a lack of strong evidence indicating a significant change in annual records in marathon or half marathon races at any given time between the 1980s and 2023 (besides the women's half marathon, which showed weak evidence for a changepoint in 1996). This stipulates that records in marathon and half marathon races will likely continue on their current trajectory, even with the addition of new technologies in the field of racing.

Although the analysis on singular changepoint models is complete, it was noted earlier that the men's half marathon plausibly contains two changepoints. This is something that will require future study and could point to additional years in which annual record trends significantly differ. Discoveries with additional changepoint models could lead to interesting considerations of both past and recent events.

Overall, based on the results from the one changepoint analysis, there is no indication that carbon

fiber-plated shoes, or recent technologies in general, have an impact on current annual record trends. The weak evidence found for a changepoint in the women's half marathon in 1996 presumably reflects the rising popularity of the race rather than a specific event or additional "trend" speculation. Further analyses must be conducted to achieve full confidence in these results across all datasets, especially if others may show aspects of a multiple changepoint model.

Part Two

Considering recent impacts in the Chicago, Boston, London, and Berlin marathons

Now that annual marathon/half marathon records have been considered for possible changepoints in the last 40 years, an additional analysis can be conducted to observe potential impacts on recent performances in major marathons. Specifically, we can consider the hypothesized improvement of times since the retail of "super shoes" was launched in 2017.

According to Gazelle Sports, a "super shoe" consists of the following [18]: A midsole super foam comprised of compressible dense foam aimed at maximizing energy return and a carbon-fiber plate that lies within the super foam, aimed at propelling the heel up and foot forward to maximize boost without expending extra energy.

Nike was the first to come out with a carbon-fiber plated shoe, publicly unveiling the Nike Vaporfly 4% as the original "super shoe" on the market. Nike's competitors followed closely, with carbon-fiber plated shoes by Adidas, Asics, Brooks, New Balance, Puma, Saucony and Hoka entering retail shortly after. However, the introduction of these shoes was not indisputable. Before the Vaporflys came out publicly, they debuted as prototypes in the 2016 Olympics, causing controversy within the running community. All three men's podium winners of the 2016 Olympic marathon wore the unreleased prototypes disguised as one of Nike's publicly available racing shoes [16]. Debates naturally followed over the legality of the shoes and their competitive advantage. Many studies since 2016 have attempted to quantify the advantage of the shoes. As such, World Athletics has put out guidelines on what shoes can and cannot be worn at major races. In January of 2020, World Athletics banned shoes that have not been on public sale for four months to mitigate controversy over prototypes [12]. In addition, World Athletics amended their shoe standards to exclude shoes with more than one rigid structure (including carbon plates), and banned 'sensing or intelligent' technology embedded within a shoe [20]. These standards reflect public speculation surrounding "super shoes," although many studies remain incomplete or statistically insufficient.

The following analysis aims to investigate the speculation surrounding "super shoes," specifically whether they could have impacted recent marathon performances.

7. Data

The data used for the following analysis was manually web scraped off a variety of websites. Data specific to the Chicago and London Marathons were scraped from Marathon Guide [13][14], data specific to the Boston Marathon was scraped off the Boston Athletic Association website [7], and data specific to

the Berlin Marathon was scraped from the Berlin Marathon website [6]. The scraped data comprises four separate datasets, one for each marathon. Thus, considerations could be made for possible years with "outlier performances," consisting of slower times, potentially due to extreme temperatures.

Initially, the four datasets were constructed to contain the top 20 performances from each year since 2009, considering the men's and women's results separately. The intent was to have similar amounts of data before and after 2017, signifying the public introduction of the Nike Vaporfly 4%. A few exceptions to the general data collection should be noted, one of which occurs in the Boston Marathon dataset. In 2018, the Boston Marathon recorded temperatures in the upper 30s with hard rain, leading to overwhelmingly slow performances across the board [15]. Thus, the year 2018 was excluded from the original dataset. In addition, the Berlin Marathon dataset excludes the year 2024, given those results are not yet official on the Berlin Marathon website. Finally, we must give regard to the year 2020, in which none of the major races occurred due to the pandemic. Given these considerations, the initial data collection was not sufficient to provide valuable information about the speculated impact of carbon fiber-plated shoes, thus additional data collection had to follow the first round of analysis.

To correct for insufficiencies in the initial data collection, each dataset was expanded to include the top 20 performances from various years spanning back to 2001. The number of years sufficient to display speculated impacts in recent marathon performances varies across datasets. Thus, the choice of year is somewhat arbitrary, given each analysis is dependent on the specified race, as they differ greatly in the course, times, and targeted running community (world record holders, charity runners, etc.). Since the number of years for each dataset was somewhat arbitrary and highly dependent on the race, trial and error was necessary to find a sufficient amount of data for each analysis. However, trial and error can be justified in this circumstance as there is not a scientific method to measure how many years would be sufficient for each race independently, especially considering the large variability between races.

The Chicago Marathon requires data extending back to 2002 for sufficient analysis, considering both the men's and women's races. The data includes 20 observations from each year; however, an exception should be noted for 2007. The Chicago Marathon recorded temperatures near 90 degrees with over 70% humidity this year, leading to unusually slow performances and a cancellation of the race approximately 3.5 hours after the start. Thus, it was removed from the respective Chicago Marathon dataset.

The London Marathon requires data extending back to 2002 for the men's race and 2004 for the women's race. Again, the data includes 20 observations from each year, and there are no noted exceptions or exclusions from the race.

The Boston Marathon requires data extending back to 2004 for the men's race and 2002 for the women's race. Each year includes 20 observations, and the exclusion of 2018 due to extreme weather holds for this dataset.

The Berlin Marathon requires data extending back to 2003 for the men's race and 2005 for the women's race. The data includes 20 observations from each year with no noteworthy exclusions or exceptions. However, the unofficial results for 2024 were added to this dataset for completeness of analysis [5].

The data is consistently structured across the four datasets with 6 features as follows: *Year*, *Race*, *Sex*, *Place*, *Time*, and *Name*. The times are recorded as hours:minutes:seconds in the raw data, but are converted to seconds for ease of interpretation and consistency throughout the analysis. Features necessary to the analysis are *Year*, *Sex*, and *Time*. It is essential the times are ordered from fastest to slowest; *Place* is irrelevant after utilized to ensure proper ordering. The datasets are initially separated by race; hence, the *Race* feature is unnecessary. The *Name* feature served to ensure accuracy in initial data collection but is superfluous in this analysis. Thus, *Race*, *Place*, and *Name* were removed to pursue the analysis. It should be noted *Sex* is not used to execute methodologies in the analysis but serves as a distinction between male and female data, as both are compiled to comprise one dataset for each race. The initial 20 observations from each year can be viewed in the figures below, considering each race independently.

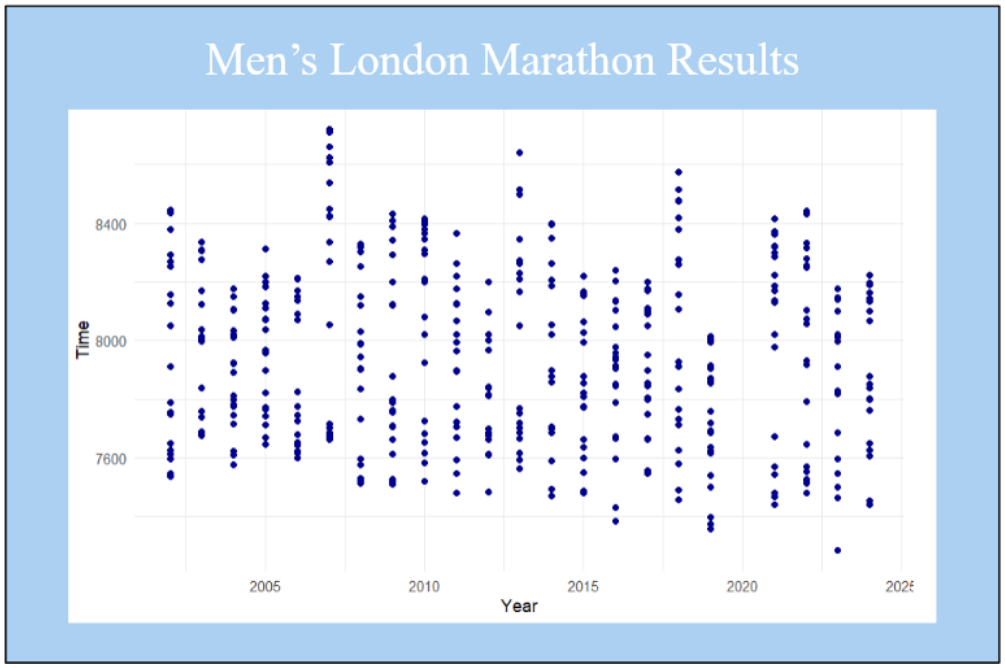


Figure 18: Top 20 performances in each year since 2002 for the men's London Marathon.

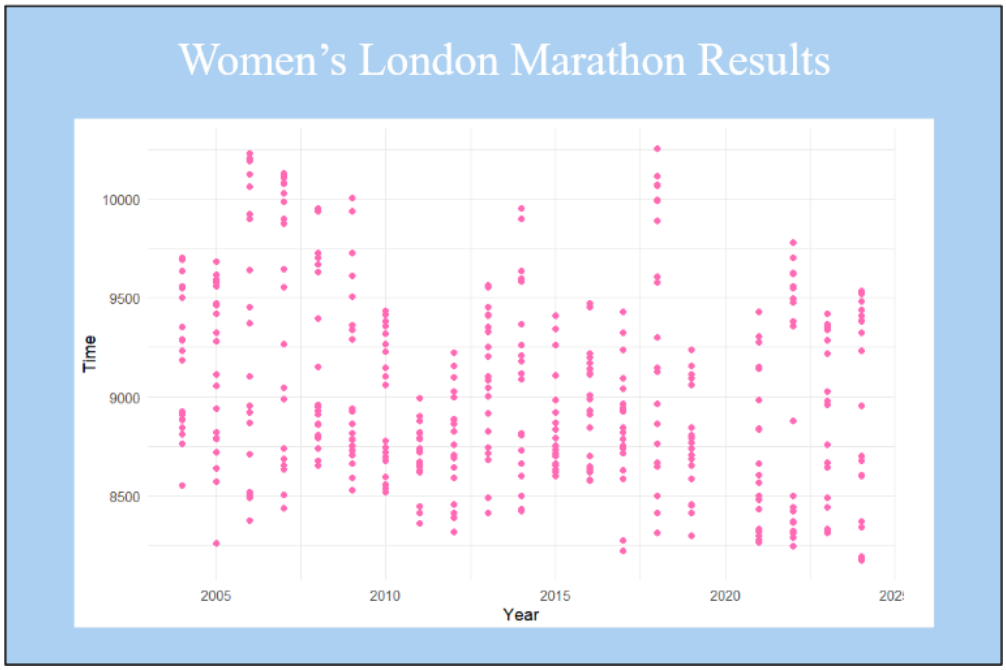


Figure 19: Top 20 performances in each year since 2004 for the women's London Marathon.

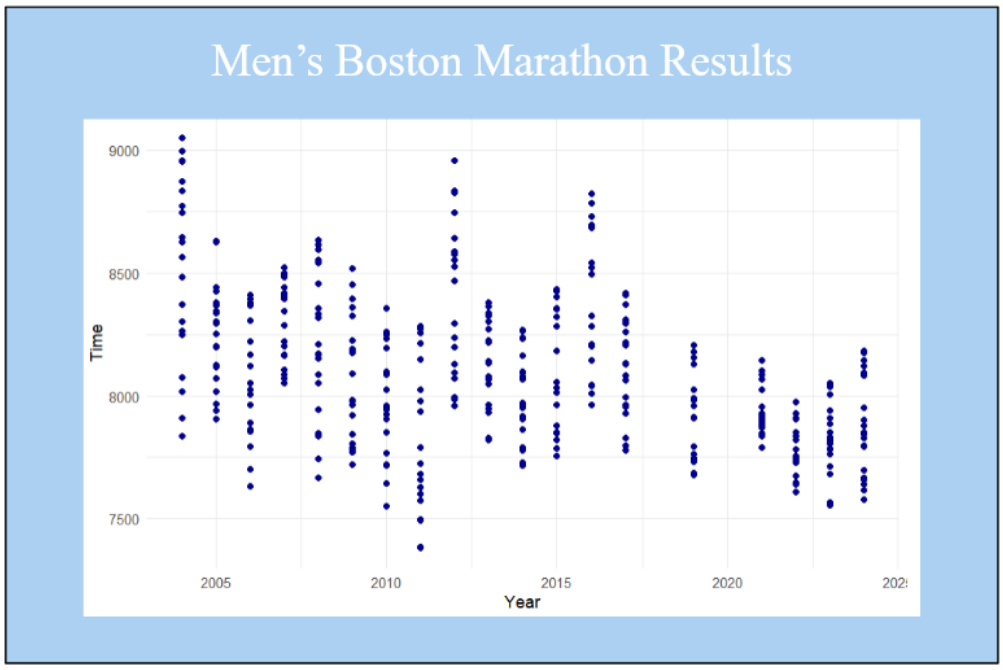


Figure 20: Top 20 performances in each year since 2004 for the men's Boston Marathon.

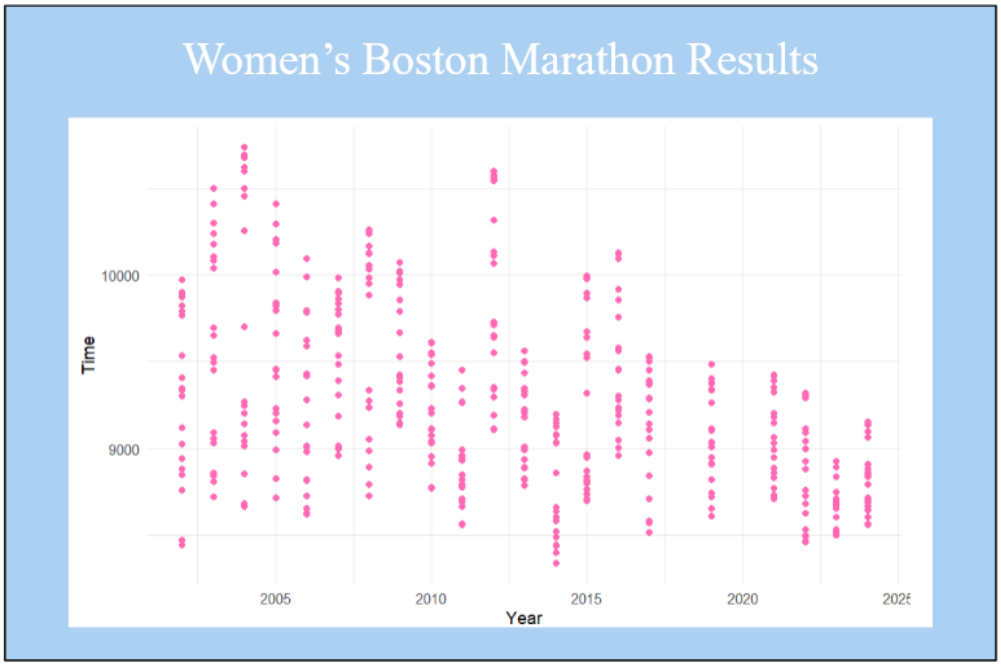


Figure 21: Top 20 performances in each year since 2002 for the women's Boston Marathon.

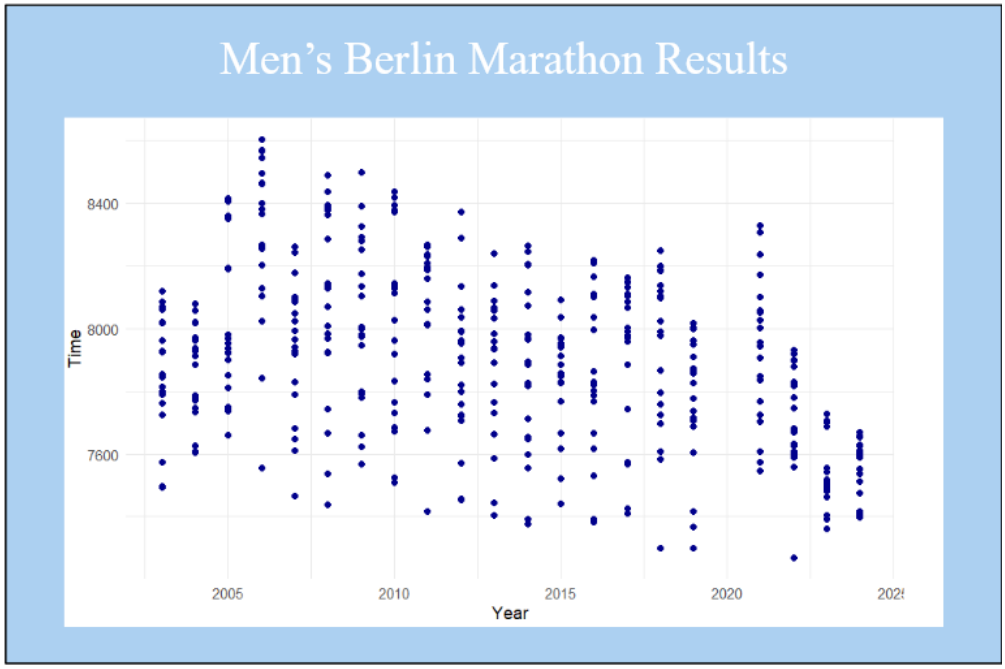


Figure 22: Top 20 performances in each year since 2003 for the men's Berlin Marathon.

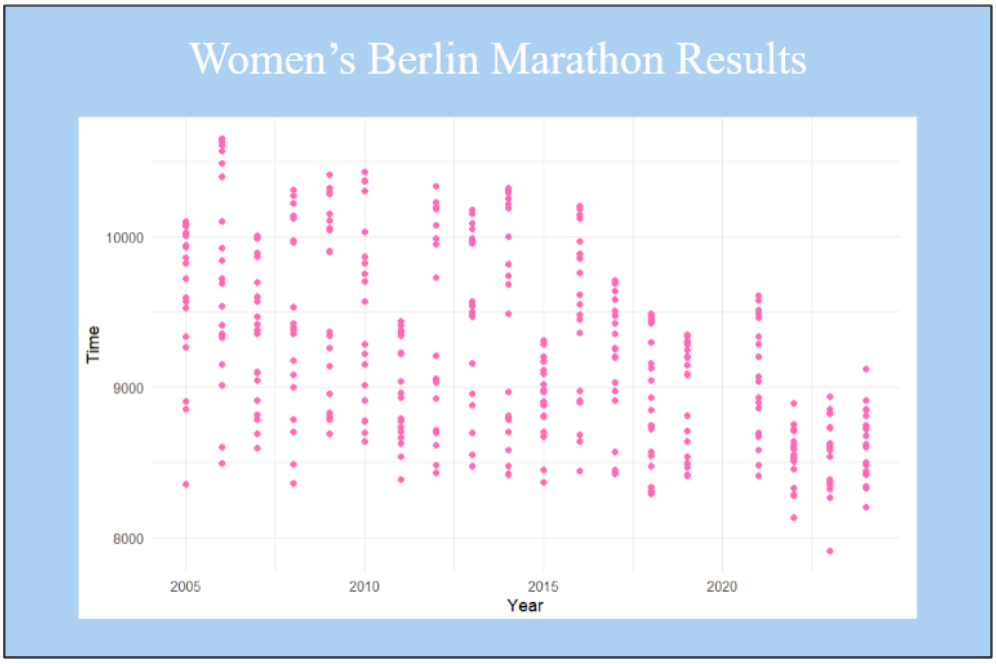


Figure 23: Top 20 performances in each year since 2005 for the women's Berlin Marathon.

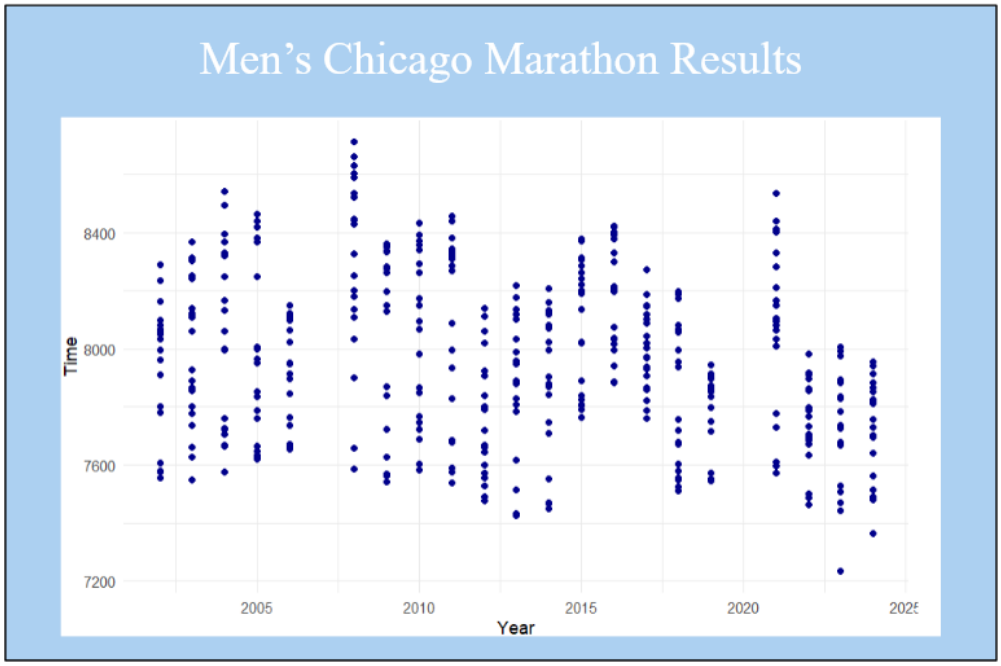


Figure 24: Top 20 performances in each year since 2001 for the men's Chicago Marathon.

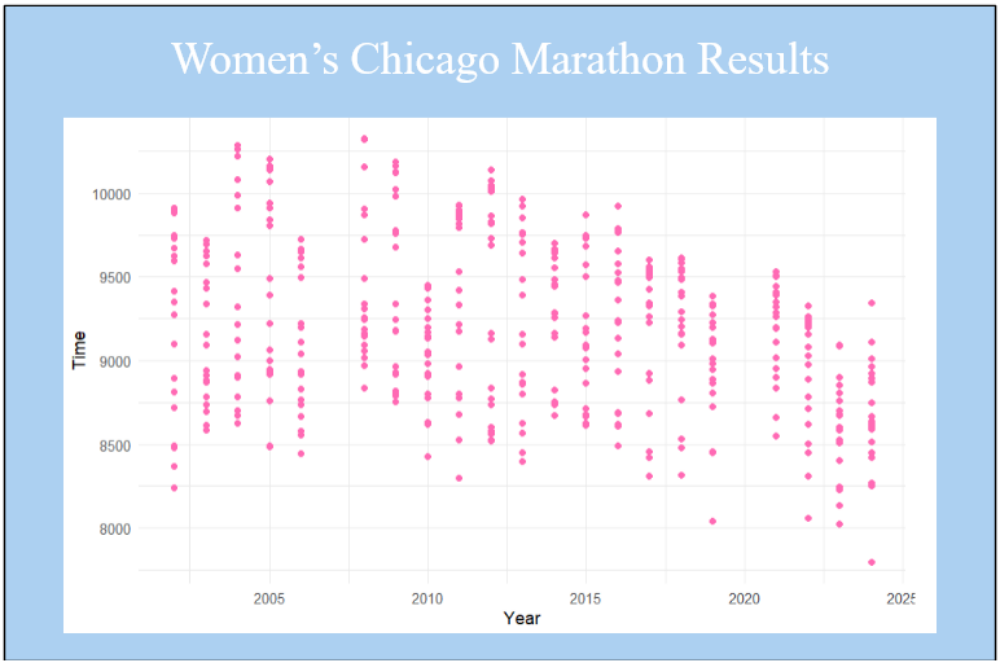


Figure 25: Top 20 performances in each year since 2001 for the women's Chicago Marathon.

8. Methods

A difficulty in implementing any extreme value analysis is the limited amount of data for model estimation. Thus, independently for each race, we can consider the fastest 20 times from each respective year, which can be loosely defined as a sequence of independent and random variables $X_1, X_2, \ldots$ that compose $M_n^{(}(k))$, where $M_n^{(k)} =$ kth largest of $X_1, \ldots, X_n$. Each $M_n^{(k)}$ represents a block, which corresponds to one year of observations. Then, the aim is to characterize the extremal behavior of X_i by examining the limiting behavior of $M_n^{(k)}$, for a fixed k, as $n \to \infty$. This is given by [8, Theorem 3.4, p. 67].

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\{(M_n - b_n)/a_n \leq z\} \to G(z) \text{ as } n \to \infty$$

for some non-degerate distribution function G, so that G is the GEV distribution function given by

$$G(z) = \exp\{-[1 + \xi(z - \mu/\sigma]^{\frac{-1}{\xi}}\},$$

then, for a fixed k,

$$Pr\{(M_n^{(k)} - b_n)/a_n \leq z\} \to G_k(z) \text{ on } \{z : 1 + \xi(z - \mu)/\sigma > 0\}, \text{where}$$

$$G_k(z) = \exp\{-\tau(z)\} \sum_{s=0}^{k-1} \frac{\tau(z)^s}{s!}$$

with

$$\tau(z) = [1 + \xi(\frac{z-\mu}{\sigma})]^{\frac{-1}{\xi}}$$

This implies if the kth largest order statistic in a block is normalized in the same way as the maximum, its limiting distribution has parameters that correspond with the parameters of the limiting GEV distribution of the block maximum.

Although this provides a limiting distribution that shares the parameters of the limiting GEV distribution for block maximum given the kth largest order statistic in a block, the model above does not include the largest r order statistics within each of several blocks, for some value r. Then, we would have to modify $M_n^{(k)}$ to be $M_n^{(r)} = (M_n^{(1)}, \ldots, M_n^{(r)})$ for each of the blocks, again each of which corresponds to one year of observations.

The necessity behind a new model is the lack of a joint distribution of $M_n^{(r)}$. In addition, the components cannot be independent, so the outcome of each component influences the distribution of the other, which can lead to further issues when analyzing the data. Thus, we have the following theorem [8, Theorem 3.5, p. 68]:

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\{(M_n - b_n)/a_n \leq z\} \to G(z) \; as \; n \to \infty$$

For some non-degenerate distribution function G, then, for fixed r, the limiting distribution as $n \to \infty$ of

$$\tilde{M}_n^{(r)} = \left(\frac{M_n^{(1)} - b_n}{a_n}, ..., \frac{M_n^{(r)} - b_n}{a_n} \right)$$

falls within the family having joint probability density function [8, Equation 3.15, p. 68]

$$f(z^{(1)}, ..., z^{(r)}) = \exp\left\{ -\left[1 + \xi \left(\frac{z^{(r)} - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}$$

$$\times \prod_{k=1}^{r} \sigma^{-1} \left[1 + \xi \left(\frac{z^{(k)} - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}-1}$$

where $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty; z^{(r)} \leq z^{(r-1)} \leq ... \leq z^{(1)}$; and $z^{(k)} : 1 + \xi(z^{(k)} - \mu)/\sigma > 0$ for $k = 1, ..., r$.

In the case where $r = 1$, the above joint probability density function is reduced to the GEV family of density functions; although $r = 1$ is not directly relevant for this analysis, it is important to note this observation to support the validity of the analysis. In addition, although we are not directly observing the case where $\xi = 0$ in the above joint probability density function, it should be noted the limiting form as $\xi \to 0$ leads to the family of density functions as follows [8, Equation 3.16, p. 68]:

$$f(z^{(1)}, ..., z^{(r)}) = \exp\left\{-\exp\left[-\left(\frac{z^{(r)} - \mu}{\sigma}\right)\right]\right\}$$

$$\times \prod_{k=1}^{r} \sigma^{-1} \exp\left[-\left(\frac{z^{(k)} - \mu}{\sigma}\right)\right]$$

For which the case $r = 1$ reduces to the density of the Gumbel family.

Now that we have established the motivation of using r-largest order statistics, and have supported the validity of modeling such data with extreme values, we can further specify how this model works with the given data.

Consider we have a series of independent and identically distributed variables, which are the times, in seconds, of each top 20 performance in a major marathon race. Then, the data is grouped into m blocks, which represents the number of years the data spans. For the data given, m is 20 for the men's and women's Chicago Marathon; although the data spans back to the year 2002 for both races, 2007 and 2020 are excluded. The men's Boston Marathon has an m of 18, given the data spans back to 2004, but excludes both 2018 and 2020. The women's Boston Marathon has an m of 20, as the data extends back to 2002, also excluding 2018 and 2020. The women's Berlin Marathon has an m of 18, given the dataset spans back to 2005, with the exclusion of 2020. For the men's Berlin Marathon, m is 20, again with the only exclusion being 2020. The men's London Marathon has an m of 21, and the women's London Marathon has an m of 19, given the men's data extends back to 2002 and the women's spans back to 2004, noting neither have exclusionary years besides 2020.

Thus, m is dependent on the dataset, which is reasonable as each dataset is analyzed separately.

Then, given m is defined, in block i the largest r_i observations are recorded, leading to the series $M_i^{(ri)} = (z_i^{(1)}, \ldots, z_i^{(r_i)})$, for $i = 1, \ldots, m$. In this analysis, $r_1 = \ldots = r_m = r = 20$, given 20 observations were available for every race in each year; if there were not 20 observations available to record, it would also be sufficient for r to vary between blocks.

The choice in r is based off a bias-variance trade-off; small values of r generates fewer data, leading to high variance, where large values of r are likely to violate the asymptotic support of the model, leading to

bias. Thus, setting $r = 20$ seems a sufficient choice to minimize variance while still supporting the original model.

The likelihood for this model, obtained from (3.15) and (3.16) above can be given as follows (when $\xi \neq 0$) [8, Equation 3.17, p. 69]:

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{m} \left(\exp\left\{ -\left[1 + \xi \left(\frac{z_i^{(r_i)} - \mu}{\sigma} \right) \right]^{-1/\xi} \right\} \times \prod_{k=1}^{r_i} \sigma^{-1} \exp\left[-\left(\frac{z_i^{(k)} - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}-1} \right)$$

Provided $1 + \xi(z^{(k)} - \mu)/\sigma > 0, k = 1, ..., r_i, i = 1, ..., m$; otherwise the likelihood is zero.

Again, in this analysis, $\xi \neq 0$, but it is important to note when $\xi = 0$, the likelihood can be given as follows [8, Equation 3.18, p. 69]:

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{m} \left(\exp\left\{ -\exp\left[-\left(\frac{z_i^{(r_i)} - \mu}{\sigma} \right) \right] \right\} \times \prod_{k=1}^{r_i} \sigma^{-1} \exp\left[-\left(\frac{z_i^{(k)} - \mu}{\sigma} \right) \right] \right)$$

It should also be noted in the case of $r_i = 1$, the likelihood function above reduces the likelihood of the GEV model for block maxima. Thus, the r largest order statistic model gives a likelihood that shares parameters with the GEV distribution of block maxima (μ, σ, ξ) while incorporating more of the observed extreme value data. This is valuable specifically to this analysis as only four races are being analyzed with a maximum of 22 years being considered; hence it requires more observations from each year to maintain an accurate model.

As we are working with running times, and given a racer's time is inverse to overall performance (the faster the time, the better the performance), the series $M_i^{(r_i)} = (zi^{(1)}, \ldots, z_i^{(r_i)})$ must be re-writen as $M_i^{(r_i)} = (-z_i^{(1)}, \ldots, -z_i^{(r_i)})$ for $i = 1, \ldots, m$. Then, the maximum likelihood estimate will be modeled as it is in (3.17), with a sign correction for $\hat{\tilde{\mu}} = -\hat{\mu}$. This will allow the estimates to account for the fact that less minutes is indicative of a faster racer, rather than the incorrect assumption of time and performance being positively correlated.

Thus, the basis of the analysis relies on the likelihood function (3.17) to perform parameter and likelihood estimations. However, for ease of interpretation, throughout this analysis the negative log of (3.17) was taken with the goal of minimizing the negative log-likelihood rather than maximizing the likelihood. It is common practice to perform such transformations with likelihood functions, as it often allows for more manageable calculations while performing various analyses.

Now that the basis and motivation of the analysis has been discussed in relation to extreme values,

the structure of the analysis can be introduced. There are three respective parts to the analyses performed independently on each dataset, described as follows:

The first is an analysis estimating the parameters μ, σ, and ξ, as well as the negative log likelihood associated with each dataset, assuming there is no trend in the data. Then, the negative log-likelihood used to calculate this estimate can be directly traced back to the likelihood function given in (3.17). The parameter estimation is computed utilizing two methods; the first of which is BFGS optimization and the other an adaptive Metropolis algorithm. Together, they provide strong evidence for the parameter estimates, with the BFGS optimization providing initial parameter estimations, negative log-likelihood, and corresponding standard errors. The adaptive Metropolis algorithm provides standard deviations and posterior distributions of the parameter estimates, as well as the information to compute a posterior distribution of the estimated endpoint for each dataset.

This analysis is repeated for a model assuming a linear trend in the data, in which everything is the same beside replacing μ with $\beta_0 + \beta_1 * i$, where i indicates the block/year, in both the negative log likelihood and subsequent parameter estimations. Then, β_0 is the mean of the recorded race times and β_1 is the estimated trend of race times from $i = 1$ to $i = m$, considering each dataset independently. Hypothetically, if race times have improved from $i = 1$ to $i = m$, β_1 should be positive, as adding β_1 to β_0 would make μ smaller, thus the average race time faster.

Finally, the described analysis is repeated for a model assuming a changepoint in 2017, which corresponds with the retail of the Nike Vaporfly 4%. This is calculated by replacing μ with $\beta_0 + \beta_1 * (i - t_0)$ for $i \leq 2016$, in which case we have $t_0 = 2016$, and replacing μ with $\beta_0 + (\beta_1 + \beta_2) * (i - t_0)$ when $i > 2016$. Here, the prediction of β_1 and β_2 is more difficult, as the degree to which mu changes depending on i also influences the optimal β_1 and β_2 values.

The two parameter estimation methods mentioned are integral to the results and interpretation of the analyses. Thus, they are described in more detail below.

The BFGS optimization algorithm [**21**] stems from gradient descent, given by the following:

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

Where α is a positive real number known as the learning rate. The equation above represents an iteration of gradient descent, or the direction in where $f(x)$ decreases the fastest, with the aim of finding a local minimum. However, this is highly dependent on the learning rate and largely inefficient, thus another

method of optimization should be considered.

Then, Newton's method can be examined, where the iterative scheme can be written as the following:

$$x_{k+1} = x_k - H(x_k)^{-1}\nabla f(x_k)$$

Where H is the hessian matrix, otherwise known as a square matrix of the second-order partial derivatives of f. This method does not require a learning rate, however it is sensitive to initial conditions and is very computationally expensive. Thus, many optimization techniques rely on quasi-Newton Methods. The difference between quasi-Newton and Newton methods is in how the Hessian matrix is computed. In quasi-Newton methods, instead of being computed, the Hessian matrix is approximated with a positive-definite matrix B, updated from iteration to iteration. However, the Hessian approximation B must satisfy the quasi-Newton condition:

$$B_{k+1}[x_{k+1} - x_k] = \nabla f(x_{k+1}) - \nabla f(x_k),$$

Letting $yk = \nabla f(x_{k+1}) - \nabla f(x_k)$ and $\Delta x_k = x_{k+1} - x_k$, so that we have

$$B_{k+1}\nabla x_k = y_k$$

Now, we can specify the type of quasi-Newton method used for this analysis, also known as the BFGS method. This method aims at uniquely specifying the Hessian estimate B in multiple dimensions by imposing additional constraints on B. This is one of the main weaknesses of the general quasi-Newton method, as it does not uniquely specify the Hessian estimate B in more than one dimension. Thus, there are many quasi-Newton methods that aim at uniquely constraining the solution.

For the BFGS method, the problem can be formulated as follows:

$$\min_{B_{k+1}} ||B_{k+1} - B_k||$$

$$\text{subject to} B_{k+1}^T = B_{k+1}$$

$$\text{and} B_{k+1}\Delta x_k = y_k.$$

Note this can be alternatively written with the approximation of B^{-1} directly. Given these constraints, the problem is equivalent to updating the approximate Hessian at each iteration by adding two symmetric, rank-one matrices U and V, where $U = auu^T$ and $V = bvv^T$, and u and v are linearly independent non-zero

vectors, given a and b are constants. This is shown by the following:

$$B_{k+1} = B_k + auu^T + bvv^T$$

Then, by imposing the quasi-Newton condition $B_{k+1}\Delta x = y_k$ and choosing $u = y_k$ and $v = B_k\Delta x_k$, we have the following:

$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T \Delta x_k} - \frac{B_k \Delta x_k \Delta x_k^T B_k}{\Delta x_k^T B_k \Delta x_k}$$

This concludes the theory behind the BFGS method. Practically, this method was used for optimization as it provided a reliable way to check if the parameter estimates were reasonable, indicated by a positive definite Hessian matrix. In addition, it provides the constraints necessary to deal with multi-dimensional data (which is a necessity given the multiple-parameter space), minimize computational stress, and maximize efficiency.

The BFGS optimization was computed by the optim() function in R, considering three, four, or five variables depending on the model (three for the no trend model, four for the linear trend model, and five for the changepoint model). This function also allows for dependency on the specified likelihood function, which was modified in a model-dependent manner, as previously specified.

Then, from this optimization method, parameter estimates were given and standard errors computed from the inverse Hessian matrix. The negative log-likelihood for each dataset was also computed from the parameter estimates provided by this method.

The second estimate utilized was the adaptive Metropolis algorithm. This method originates from a popular method used in Bayesian statistics, called the *Metropolis-Hastings algorithm*. Both algorithms utilize random numbers to simulate a sample from the posterior distribution. Computationally known as pseudorandom number generators, these methods obtain streams of numbers that look like independent, identically distributed random numbers over $(0, 1)$, where a variety of transformation techniques can convert these uniform random numbers to any desired distribution. Particularly, the Metropolis-Hastings algorithm observes the following steps (considering a discrete state space, although the continuous state space also closely reflects the following) [22, p. 45]:

Start from an arbitrary $X^{(0)}$, where X is a parameter of the model, and $X^{(0)}$ is typically taken as the maximum likelihood estimator.

Given $X^{(n)} = \xi$, generate a random variable $Y = x_J$, where J is chosen by $Pr\{J = j|X^{(n)} = x_i\} = q_{ij}$ for some family $\{q_{ij}\}$ such that the $\sum_j q_{ij} = 1$ for each i. Then, Y is the trial value for the next step of the

Markov chain.

Define $\alpha = min\left(\frac{f_j q_{ji}}{f_i q_{ij}}, 1\right)$. If $\alpha = 1$, then set $X^{(n+1)} = Y$, in which Y is accepted. However, if $0 < \alpha < 1$, Y is then accepted with probability α, thus $X^{(n+1)} = Y$. Otherwise, we have $X^{(n+1)} = X^{(n)}$.

The adaptive Metropolis algorithm is very similar to the above Metropolis-Hastings algorithm, excluding some key distinctions. An adaptive Metropolis algorithm adapts continuously to the target distribution, affecting both the size and spatial orientation of the proposal distribution. In addition, the covariance of the proposal distribution is calculated using all the previous states. The advantage of such an approach is that it starts accumulating information right at the beginning of the simulation, promoting rapid adaption and minimizing the number of function evaluations needed [11].

Mathematically, the adaptive Metropolis algorithm can be written as the following:

Suppose at time $t-1$ we have sampled the states $X_0, X_1, \ldots, X_{t-1}$, where X_0 is the initial state (see above similarities). Then, a trial value Y is sampled from the proposal distribution $q_t(.|X_0, \ldots, X_{t-1})$, which now depends on the whole history of $(X_0, X_1, \ldots, X_{t-1})$. Thus, Y is accepted with probability $\alpha(X_{t-1}, Y) = \min\left(1, \frac{\pi(Y)}{\pi(X_{t-1})}\right)$, given the target distribution has the density $\pi(x)$ on the subset S in $\mathbb{R}^d$. If Y is accepted, $X_t = Y$, otherwise $X_t = X_{t-1}$.

Then, although the two algorithms follow similar steps, the adaptive Metropolis algorithm is ideal for this analysis given its ability to rely on previous states to continuously adapt to the target distribution. Thus, we can use the parameter estimations for maximum likelihood, found in the BFGS optimization algorithm, to define the initial state X_0. In addition, the covariance matrix derived from the BFGS optimization algorithm can be of use when considering the initial covariance C_0. There are a few additional requirements to implement this algorithm into an R code. This includes a negative log likelihood function (again, this will change based on the model considered), the total number of simulations (*npar*), the initial run with the given C_0 ($n1$), an interval between subsequent updates ($n2$), an interval between updates of output parameter matrix ($n3$), and a scaling parameter (*scal*).

For this analysis, *npar* was chosen to be $10,000$. Then, $n1$ was chosen to be 500, $n2$ 200, and $n3$ 50. The initial run serves as a burn-in [22, p. 46], or a period during which the state probabilities are assumed to be settling to the true $f(.)$. The reason the burn-in period was not chosen to be very large in this analysis is because we were fairly confident in the covariance values given by the optim() function. In addition, there are few parameters to consider overall, with a max of 5, thus $n1$ does not necessarily have to be large to serve its purpose. Then, the value of $n2$ was chosen to maximize the proposal distribution's adaption to the target distribution without overfitting the proposal distribution to local characteristics. Finally, $n3$

was chosen to display enough values to accurately portray the posterior distribution without establishing accidental dependencies between the values.

Although the values above followed reasonably along model analyses, the scaling parameter utilized for the analysis strayed from the typical standard. The basic choice for a scaling parameter [11] is adopted to be 2.4. However, this led to an acceptance rate that was entirely too low to lead to sufficient analysis. Thus, *scal* was set to 0.86 to ensure enough trial values were accepted throughout the algorithm. This could have contributed to a less efficient analysis, but reducing the *scal* parameter was necessary to ensure acceptances stayed in an acceptable range, which for this analysis can be observed between 10% and 30%.

With these values, the algorithm utilized in R follows closely to the mathematical formulation of the adaptive Metropolis algorithm displayed above. The output of this model includes the posterior distribution of each parameter. In addition, standard deviations were calculated for each parameter to compare with the standard errors computed from the optim() function. Utilizing the output from the adaptive Metropolis algorithm, the estimated endpoints of each analysis were able to be computed as well. These can provide valuable information on the performance of the model as well as the estimated upper end-point of race times. The endpoints are calculated after a burn-in period of 20% (discarding the first 20% of the MCMC output), and utilizing the following equations:

$$\text{considering the no trend model: } \mu - (\sigma/\xi)$$

$$\text{considering the linear trend model: } (\beta_0 + \beta_1 * (m - n)) - (\sigma/\xi),$$

where m is the index of the most recent year (2024) and n is the index of the changepoint

$$\text{considering the changepoint model: } (\beta_0 + (\beta_1 + \beta_2) * (m - n)) - (\sigma/\xi),$$

where m is the index of the most recent year (2024) and n is the index of the changepoint.

Finally, to determine significance between the three models for each dataset, a likelihood ratio test must be performed between them. Supposing L_1 and L_2 are the negative log likelihoods of either the no trend and linear trend models, or the linear trend and changepoint models respectively, the following can be used to observe if the subsequent model is significant.

$$D = 2 * ((L_1) - (L_2)),$$

where L_1 is necessarily greater than L_2 given the optimization over a larger parameter space.

Then, a chi-squared test can be performed to determine whether the models are significant at the 0.05 level; this is utilized in conjunction with a degree of freedom = 1, since the models differ by one parameter. If the comparison is between the no trend and linear trend models, the $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$. Then, if the comparison is between the linear trend model and the changepoint model, the $H_0 : \beta_2 = 0$ and $H_a : \beta_2 \neq 0$.

It should be noted that given the location of the changepoint is considered fixed for this analysis, the χ^2 approximation is valid, in contrast with the methodology from Part 1. In addition, years and times, recorded in seconds, were divided by $1,000$ to promote numerical stability in the above methodology. The interpretation of results should necessarily reflect this.

9. Results

As briefly mentioned in the data section, the methodologies described above were originally computed on the top 20 performances from each year spanning back to 2009, regardless of the specified race.

The first observation to make from this group of data is the significance of the men's Boston Marathon changepoint model when compared to the linear trend model. However, the linear trend model itself was not significant compared to the no trend model. This should raise suspicion within the analysis, as the original data suggests a strong possibility of improved running times throughout the years provided. The tables below support this suspicion, possibly indicating the insufficient amount of data utilized for the analysis:

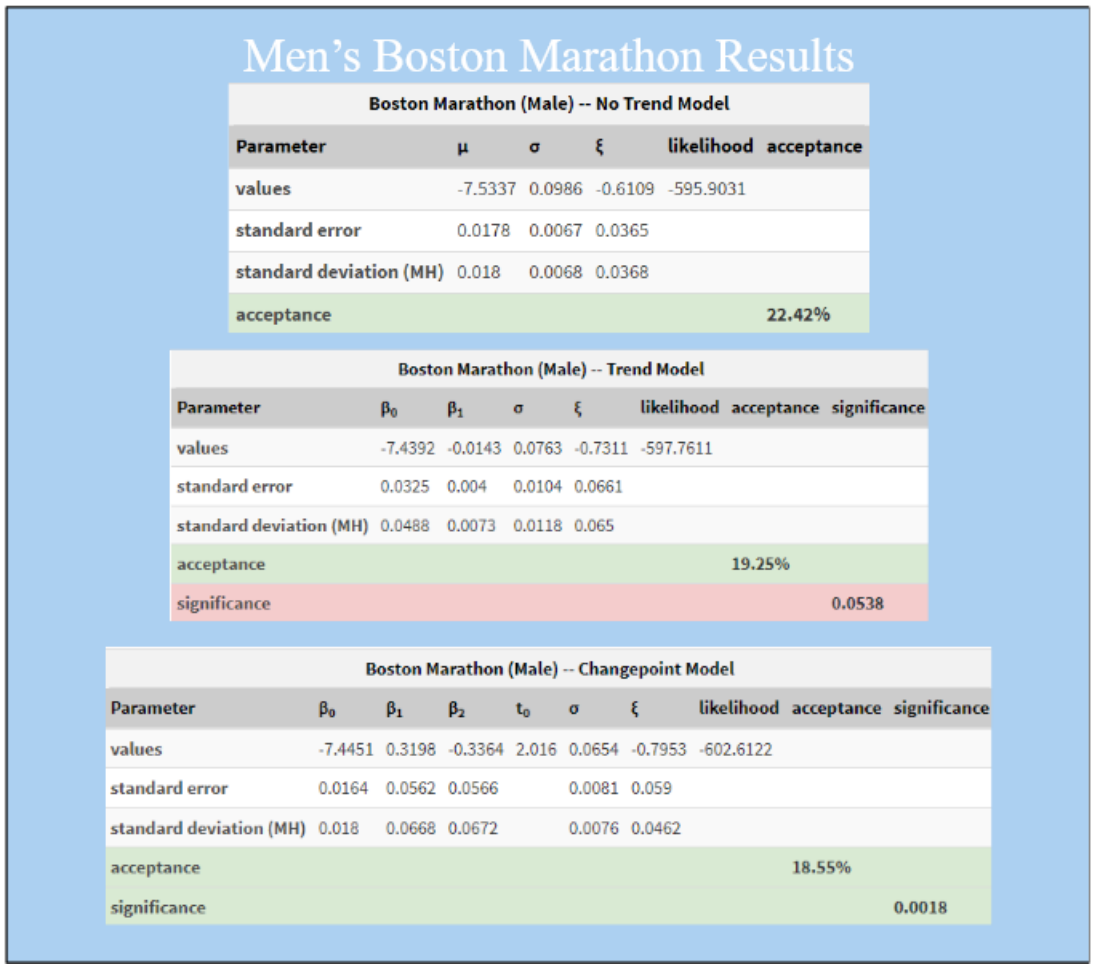


Figure 26: The estimated parameters, likelihood, corresponding standard errors/deviations, adaptive Metropolis algorithm acceptance, and significance of the liklihood ratio tests for the men's Boston Marathon. Note the significance in the linear trend table corresponds to the liklihood ratio test statistic comparing the no trend model to the linear trend model, and the significance in the changepoint table corresponds to the liklihood ratio test statistic comparing the linear trend model to the changepoint model.

Following the suspicion surrounding the men's Boston Marathon models, the women's Boston Marathon had issues with optimizing parameters; the BFGS optimization led to an immediate decrease in ξ until it was out of reasonable bounds (< -1). This indicates the possibility of a global minimum occurring at a boundary of the data, which also proved to be a problem in the adaptive Metropolis algorithm with the following trace plot of ξ:

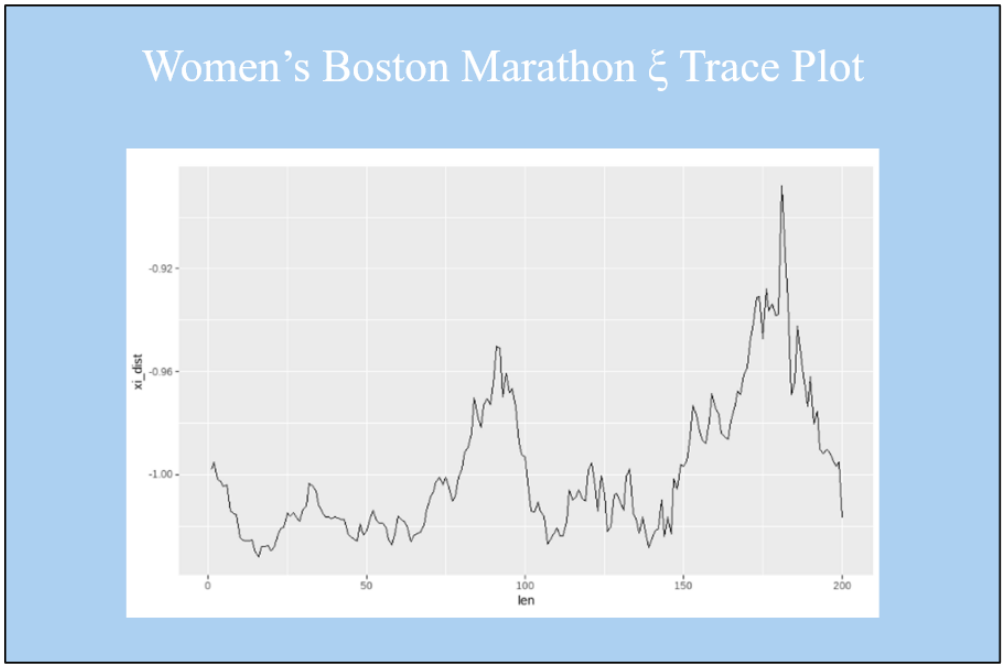


Figure 27: Women's Boston Marathon ξ trace plot

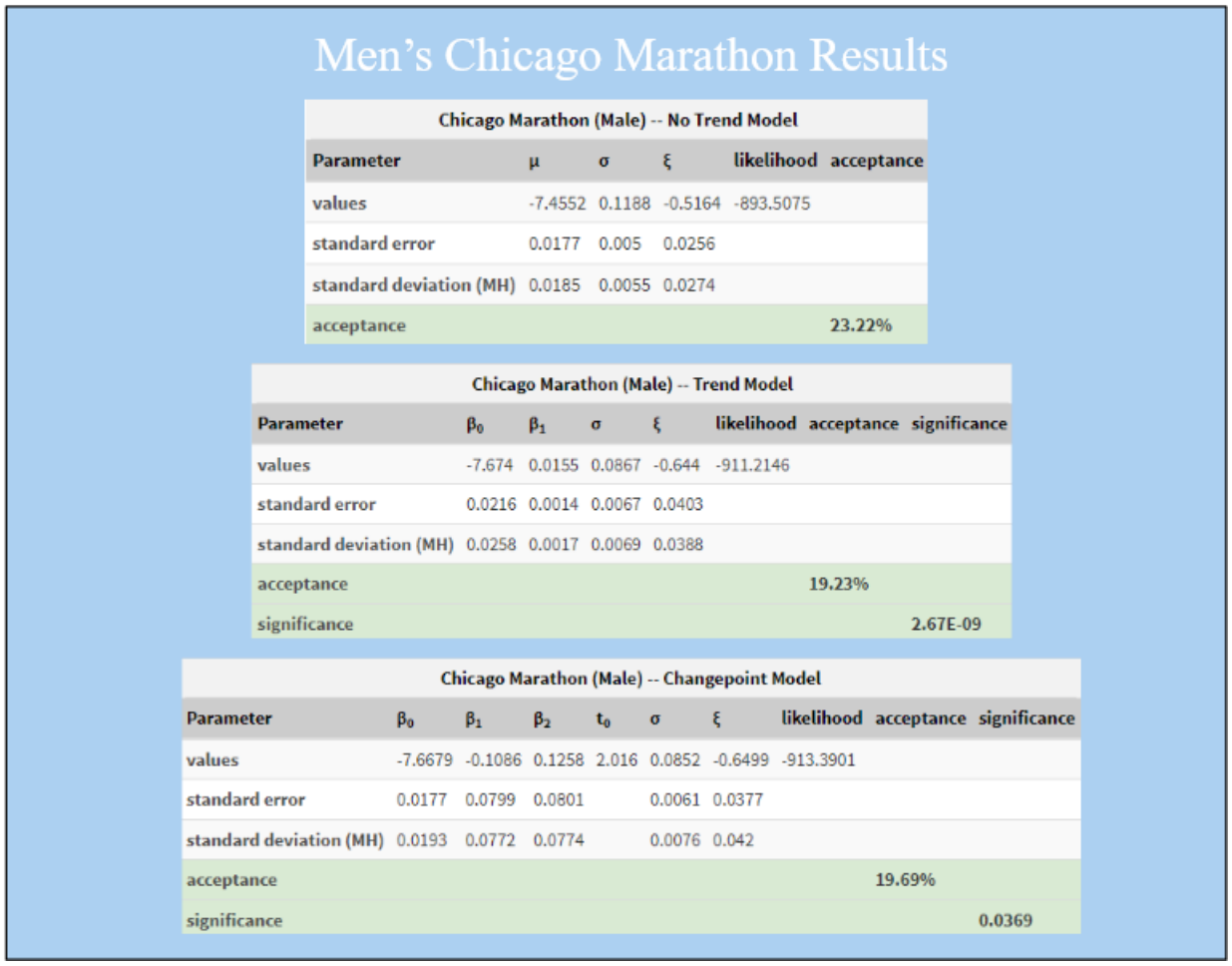
Ideally, the trace plot of the parameter estimation of ξ should display even peaks and dips that are reasonably close together. However, given this trace plot displays obvious peaks at greater intervals, we can infer there is an issue with the parameter estimation in this analysis.

These two suspicious results cautioned the remaining analysis spanning back to 2009. The other race results were relatively consistent, showing significance with the linear trend model when compared to the no trend model, but not showing significance between the changepoint and linear trend models. This means data regarding the London, Chicago, and Berlin marathons spanning back to 2009 is not sufficient to show there is anything that could have impacted recent marathon performances, outside of a linear trend that spans the entirety of the data. Again, this result was suspicious, as many of the original plots of the data displayed reasonable evidence for a changepoint.

It should also be noted that the women's London Marathon had similar issues with optimization as the women's Boston Marathon, again possibly due to a global minimum occurring at the boundary of the data. Then, the data utilized for each race had to be re-evaluated for sufficiency and accuracy within the analysis.

The revised men's Chicago Marathon data achieved the following parameter estimations and corresponding likelihood, standard errors given by the BFGS optimization, and standard deviations as given by the adaptive Metropolis algorithm (Note the significance in the linear trend table corresponds to the likeli-

hood ratio test statistic comparing the no trend model to the linear trend model, and the significance in the changepoint table corresponds to the likelihood ratio test statistic comparing the linear trend model to the changepoint model):



Chicago Marathon (Male) -- No Trend Model

Parameter	μ	σ	ξ	likelihood	acceptance
values	-7.4552	0.1188	-0.5164	-893.5075	
standard error	0.0177	0.005	0.0256		
standard deviation (MH)	0.0185	0.0055	0.0274		
acceptance					23.22%

Chicago Marathon (Male) -- Trend Model

Parameter	β_0	β_1	σ	ξ	likelihood	acceptance	significance
values	-7.674	0.0155	0.0867	-0.644	-911.2146		
standard error	0.0216	0.0014	0.0067	0.0403			
standard deviation (MH)	0.0258	0.0017	0.0069	0.0388			
acceptance						19.23%	
significance							2.67E-09

Chicago Marathon (Male) -- Changepoint Model

Parameter	β_0	β_1	β_2	t_0	σ	ξ	likelihood	acceptance	significance
values	-7.6679	-0.1086	0.1258	2.016	0.0852	-0.6499	-913.3901		
standard error	0.0177	0.0799	0.0801		0.0061	0.0377			
standard deviation (MH)	0.0193	0.0772	0.0774		0.0076	0.042			
acceptance								19.69%	
significance									0.0369

Figure 28: The estimated parameters, likelihood, corresponding standard errors/deviations, adaptive Metropolis algorithm acceptance, and significance of the liklihood ratio tests for the models representing the men's Chicago Marathon.

Note the standard errors and standard deviations closely reflect each other, indicating successful parameter estimation. Also note the acceptance for each model is around or above 20%, which is an indication of a successful adaptive Metropolis algorithm and comprehensive posterior distributions of the parameters. In addition, the linear trend model is significant when compared to the no trend model, and the changepoint model is significant when compared to the linear trend model. Perhaps the most interesting result to look at is the behavior of the endpoints as calculated from the posterior distributions provided by the adaptive Metropolis algorithm. As viewed below, the model assuming no trend has an endpoint centered around -7.212, or approximately 2 hours, 12 seconds, given the individual times were negated for the purpose of

the negative log likelihood estimation, as well as transformed to seconds and scaled down by an order of 1000. For clarification, when referring to an endpoint "centered" around a value, we are referring to the posterior mean of the calculated endpoints. It should be noted that each endpoint estimation represents the upper end of the distribution of running times, specifically for 2024. Then, for the linear trend model, we can notice an endpoint centered around -7.424, or 2 hours, 3 minutes, and 44 seconds. In addition, the changepoint model computed an endpoint posterior distribution centered around -7.414, or approximately 2 hours, 3 minutes and 34 seconds. Each of these estimated endpoint densities is also noticeably skewed, suggesting faster running times for the year 2024 are possible, although more probable estimations are found on the slower side of the density plot.

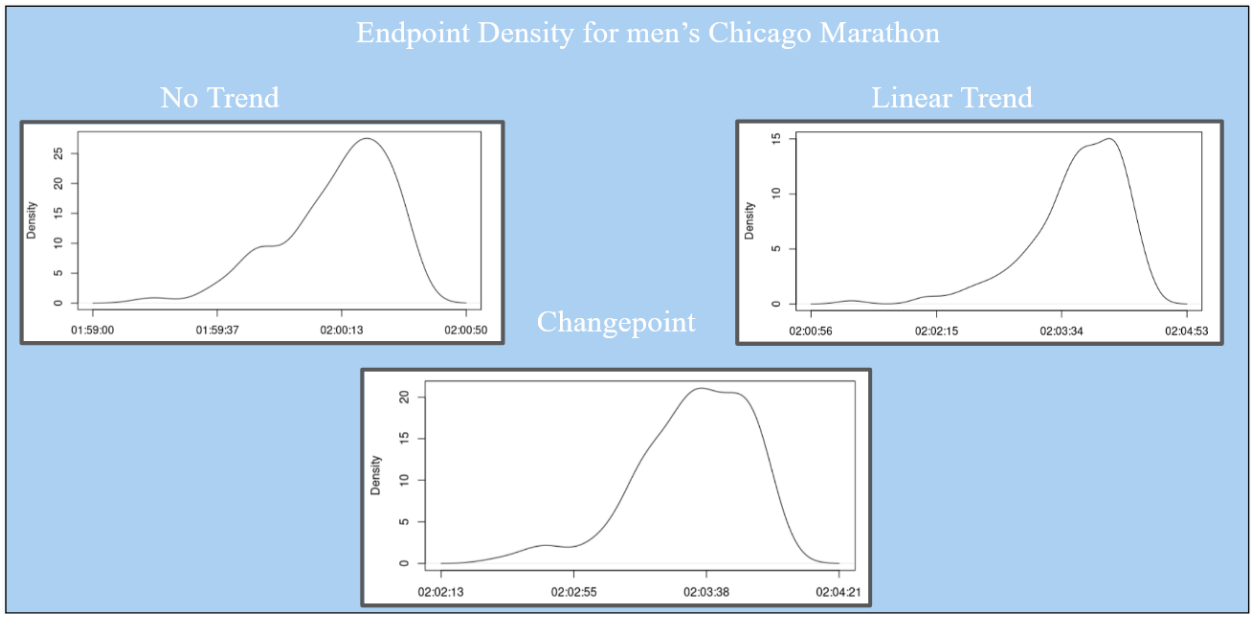
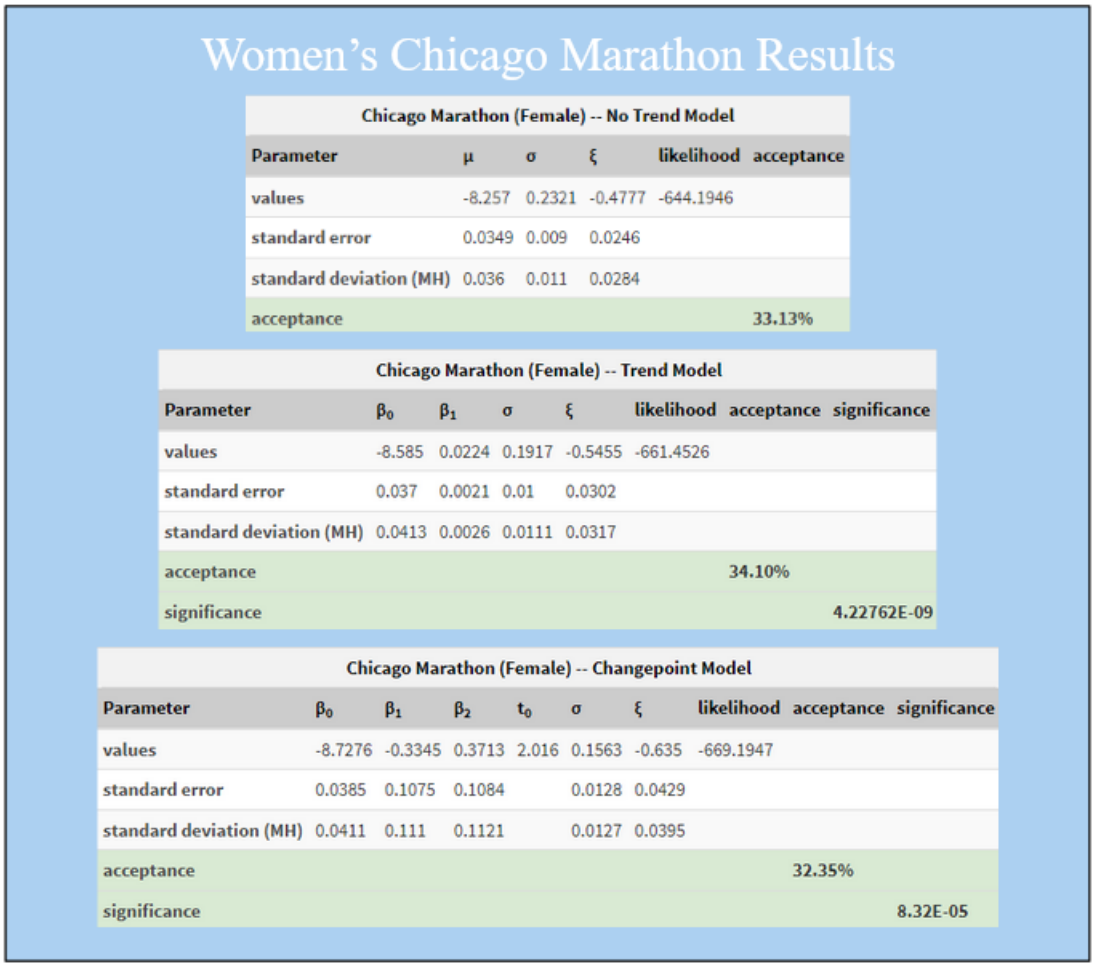


Figure 29: Density plots of the estimated endpoints, calculated from the output of the adaptive Metropolis algorithm for the men's Chicago Marathon.

Similarly, the women's Chicago Marathon, as well as both the men's and women's Boston Marathon, Berlin Marathon, and London Marathon show significance with both the linear trend model and the changepoint model, as given in the below figures. Note for each table, the acceptance rate for the adaptive Metropolis algorithm is within the range of $10\% - 35\%$, which is sufficient for this analysis. It can be observed that the women's London Marathon shows the lowest acceptance rate, which could note possible issues with the data collection and sufficiency for that particular race. Considering this, overall, the standard errors of the estimated parameters roughly correspond to the standard deviations computed from the adaptive Metropolis algorithm, indicating successful parameter and corresponding likelihood estimations. Thus, we can assume the likelihood ratio test calculated for each of these models is a reliable measure of their resulting significance.

When observing the figures below, note the significance in the linear trend tables correspond to the likelihood ratio test statistic comparing the respective no trend model to the linear trend model, and the significance in the changepoint tables correspond to the likelihood ratio test statistic comparing the respective linear trend model to the changepoint model.



Chicago Marathon (Female) -- No Trend Model

Parameter	μ	σ	ξ	likelihood	acceptance
values	-8.257	0.2321	-0.4777	-644.1946	
standard error	0.0349	0.009	0.0246		
standard deviation (MH)	0.036	0.011	0.0284		
acceptance					33.13%

Chicago Marathon (Female) -- Trend Model

Parameter	β_0	β_1	σ	ξ	likelihood	acceptance	significance
values	-8.585	0.0224	0.1917	-0.5455	-661.4526		
standard error	0.037	0.0021	0.01	0.0302			
standard deviation (MH)	0.0413	0.0026	0.0111	0.0317			
acceptance						34.10%	
significance							4.22762E-09

Chicago Marathon (Female) -- Changepoint Model

Parameter	β_0	β_1	β_2	t_0	σ	ξ	likelihood	acceptance	significance
values	-8.7276	-0.3345	0.3713	2.016	0.1563	-0.635	-669.1947		
standard error	0.0385	0.1075	0.1084		0.0128	0.0429			
standard deviation (MH)	0.0411	0.111	0.1121		0.0127	0.0395			
acceptance								32.35%	
significance									8.32E-05

Figure 30: The estimated parameters, likelihood, corresponding standard errors/deviations, adaptive Metropolis algorithm acceptance, and significance of the liklihood ratio tests for the models representing the women's Chicago Marathon.

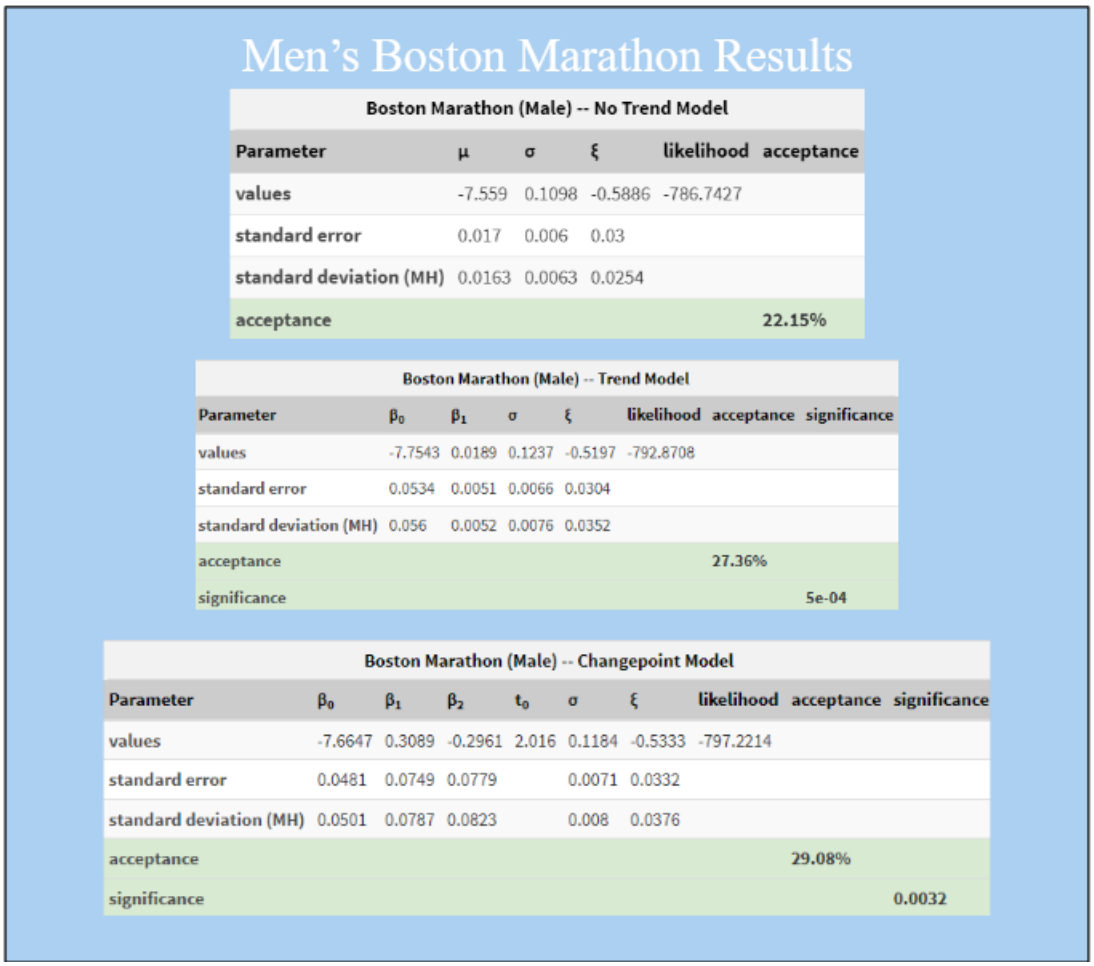


Figure 31: The estimated parameters, likelihood, corresponding standard errors/deviations, adaptive Metropolis algorithm acceptance, and significance of the liklihood ratio tests for the models representing the men's Boston Marathon.

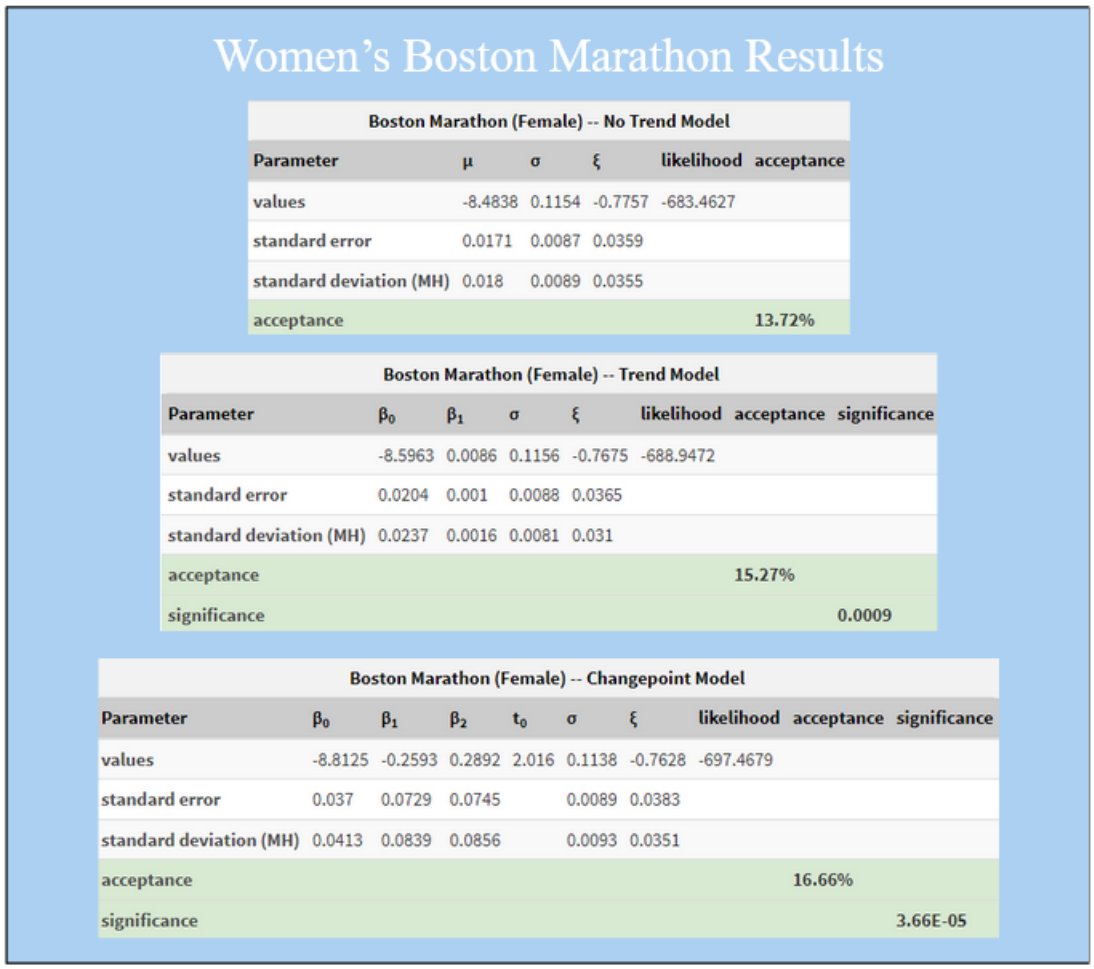


Figure 32: The estimated parameters, likelihood, corresponding standard errors/deviations, adaptive Metropolis algorithm acceptance, and significance of the liklihood ratio tests for the models representing the women's Boston Marathon.

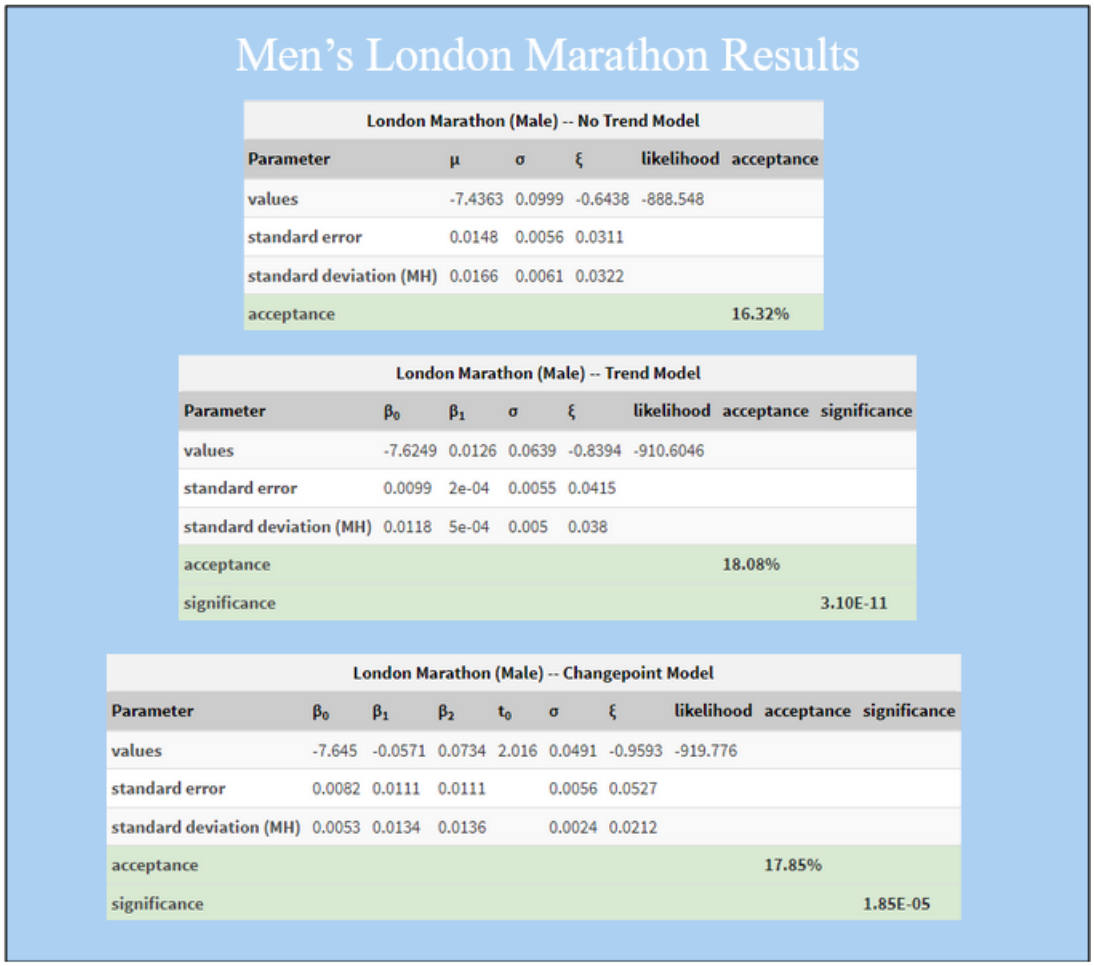


Figure 33: The estimated parameters, likelihood, corresponding standard errors/deviations, adaptive Metropolis algorithm acceptance, and significance of the liklihood ratio tests for the models representing the men's London Marathon.

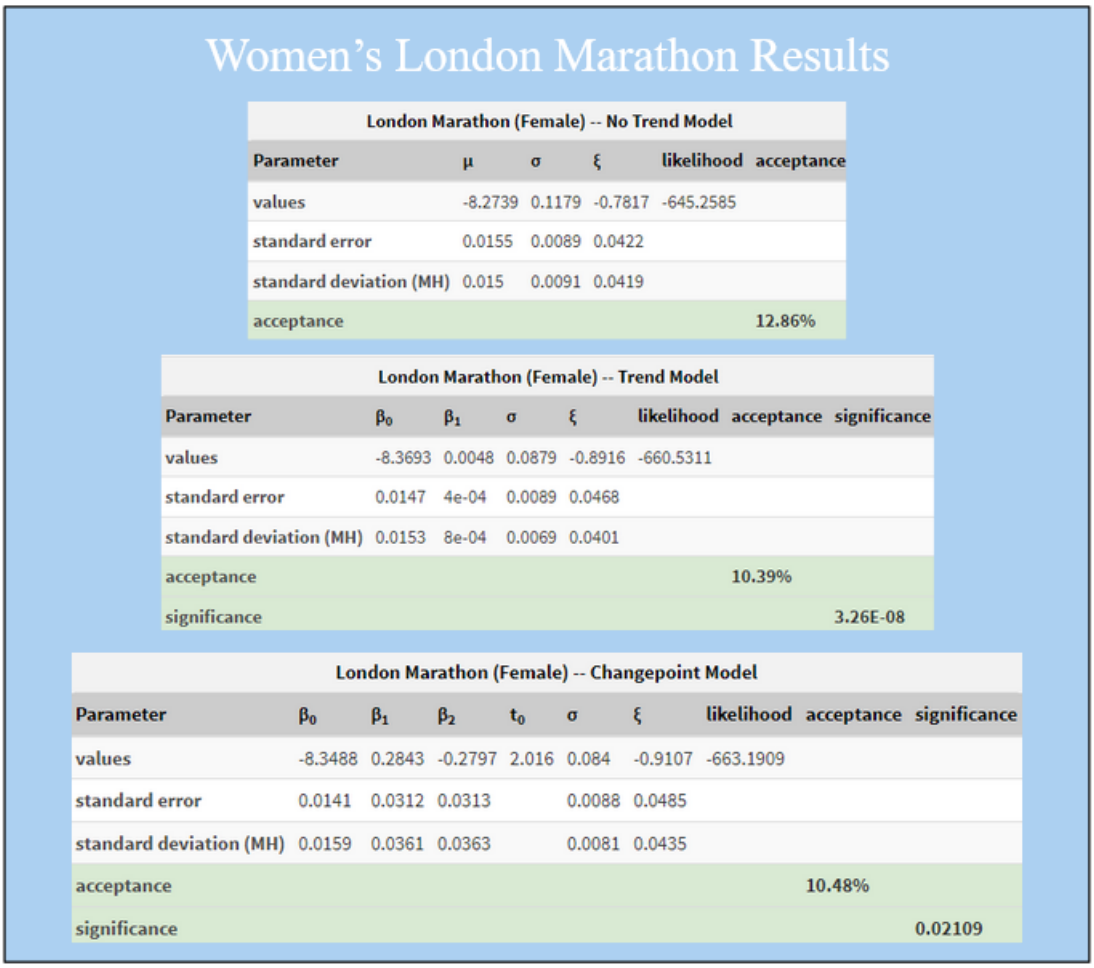


Figure 34: The estimated parameters, likelihood, corresponding standard errors/deviations, adaptive Metropolis algorithm acceptance, and significance of the liklihood ratio tests for the models representing the women's London Marathon.

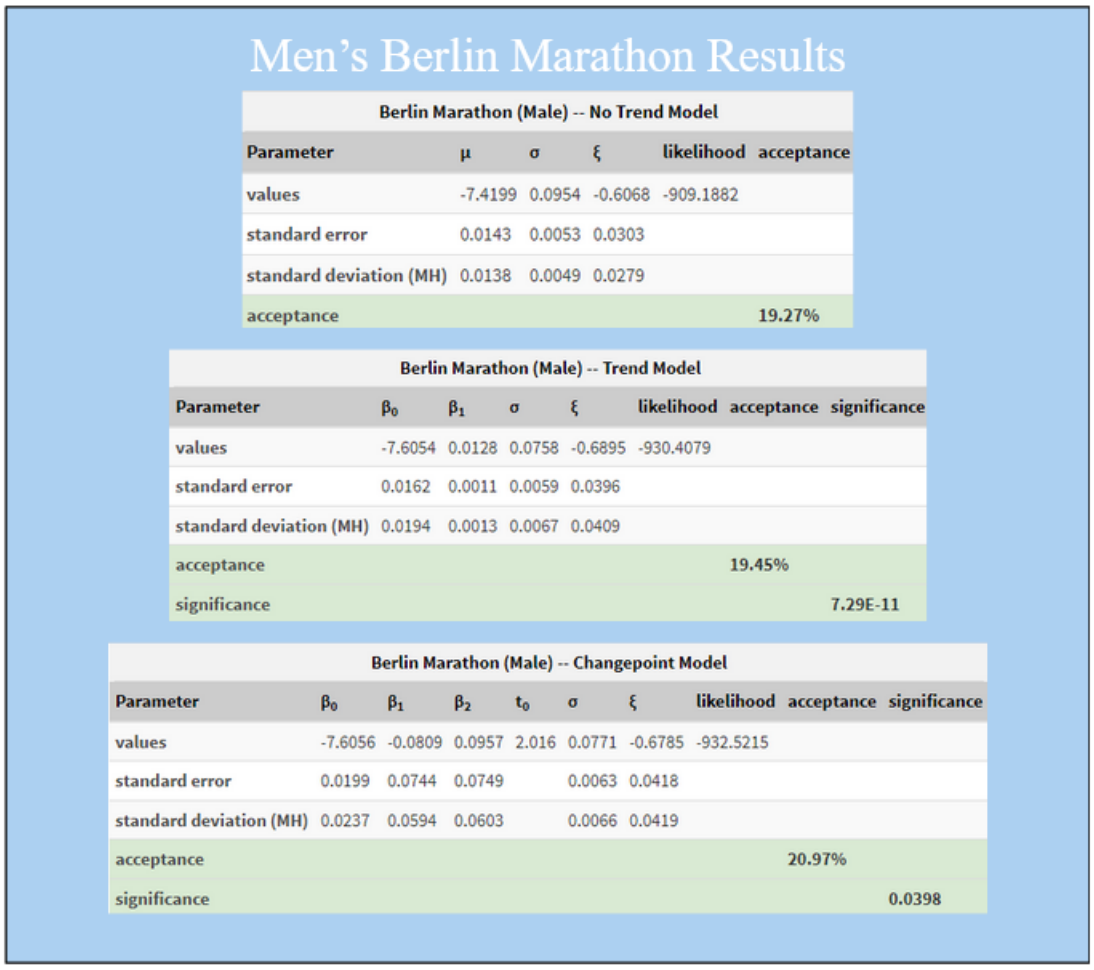


Figure 35: The estimated parameters, likelihood, corresponding standard errors/deviations, adaptive Metropolis algorithm acceptance, and significance of the liklihood ratio tests for the models representing the men's Berlin Marathon.

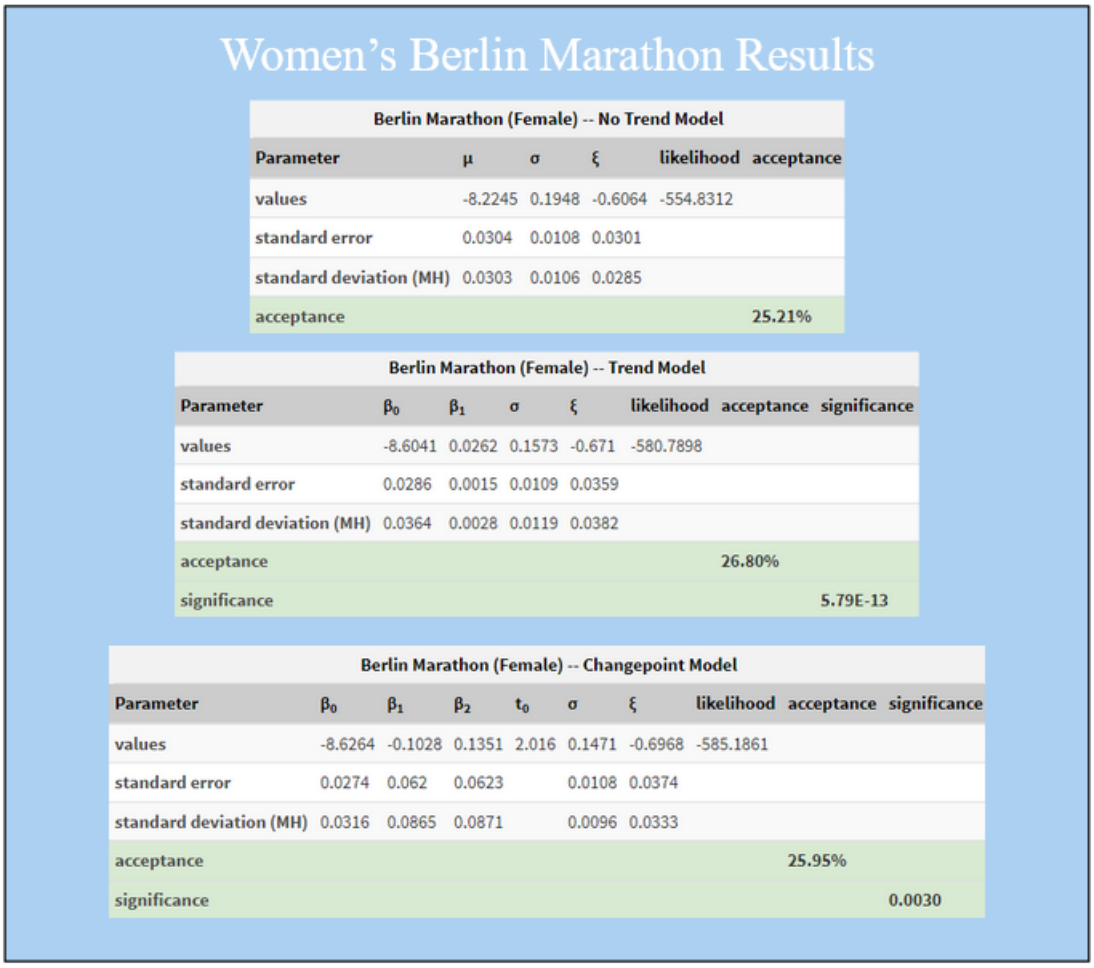


Figure 36: The estimated parameters, likelihood, corresponding standard errors/deviations, adaptive Metropolis algorithm acceptance, and significance of the liklihood ratio tests for the models representing the women's Berlin Marathon.

Examining the rest of the endpoint calculations from each respective adaptive Metropolis analysis, we see the women's Chicago Marathon estimates endpoint times centered around 2 hours, 12 minutes, and 57 seconds for the no trend model. Considering the linear trend model, the women's Chicago Marathon estimates endpoint times with a posterior mean of 2 hours, 15 minutes, and 22 seconds. The changepoint model estimates endpoint times centered around 2 hours, 17 minutes, and 43 seconds. The corresponding figures can be viewed below; again, note the right-skewed distribution of each density plot corresponding to the possibility of perceiving faster times than anticipated.

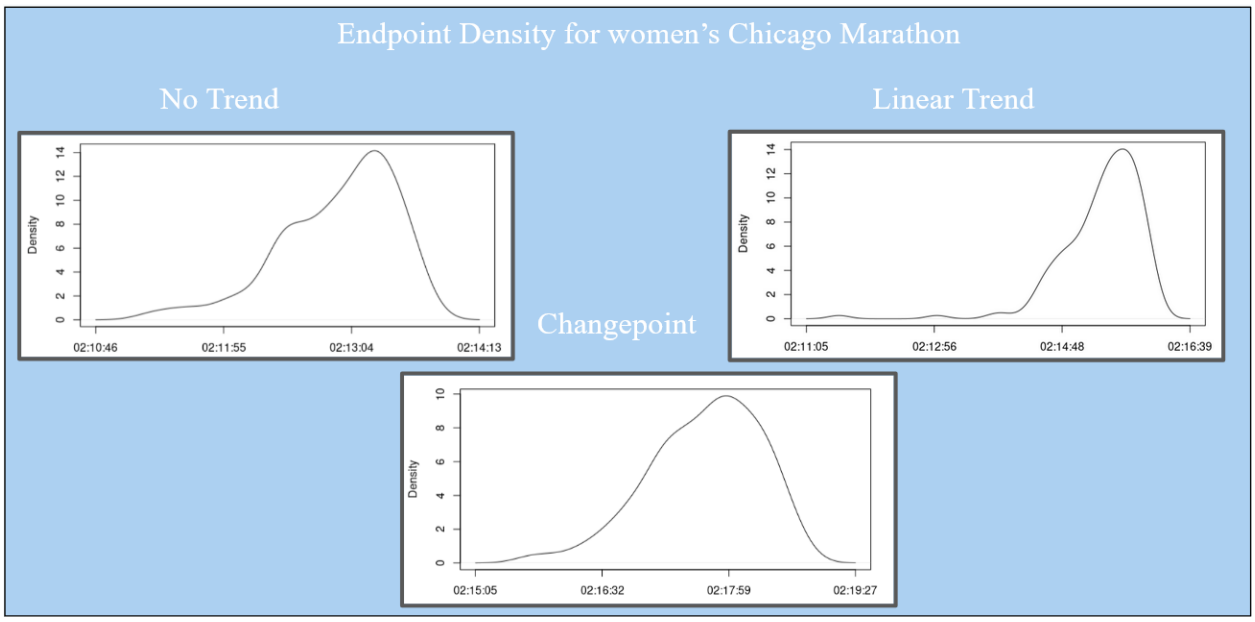


Figure 37: Density plots of the estimated endpoints, calculated from the output of the adaptive Metropolis algorithm for the women's Chicago Marathon.

Then, the men's London Marathon estimates endpoint times centered around 2 hours, 1 minute, and 12 seconds for the no trend model. The linear trend model projects endpoint times around 2 hours, 4 minutes, and 25 seconds. Considering the changepoint model, the men's London Marathon projects endpoint times around 2 hours, 4 minutes, and 51 seconds. These results can be viewed in the figures below.

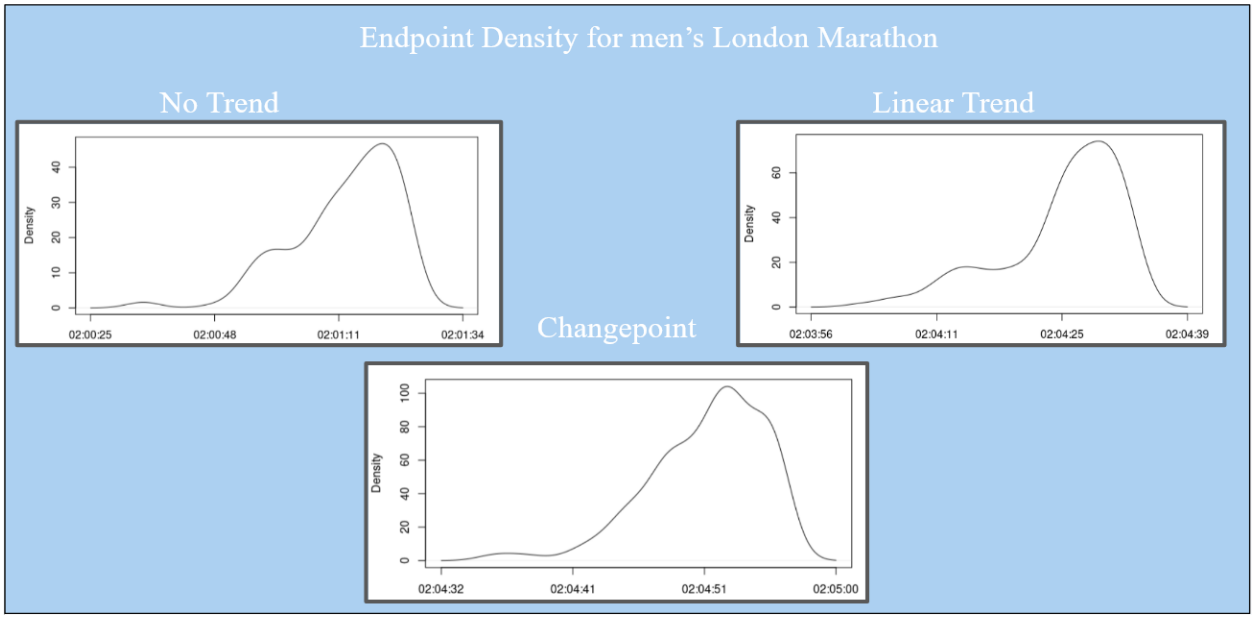


Figure 38: Density plots of the estimated endpoints, calculated from the output of the adaptive Metropolis algorithm for the men's London Marathon.

The women's London Marathon estimates endpoint times centered around 2 hours, 16 minutes, and 55 seconds for the no trend model. The linear trend model projects endpoint times around 2 hours, 17 minutes, and 15 seconds. Considering the changepoint model, the women's London Marathon projects endpoint times around 2 hours, 17 minutes, and 07 seconds. These results can be viewed in the figures below.

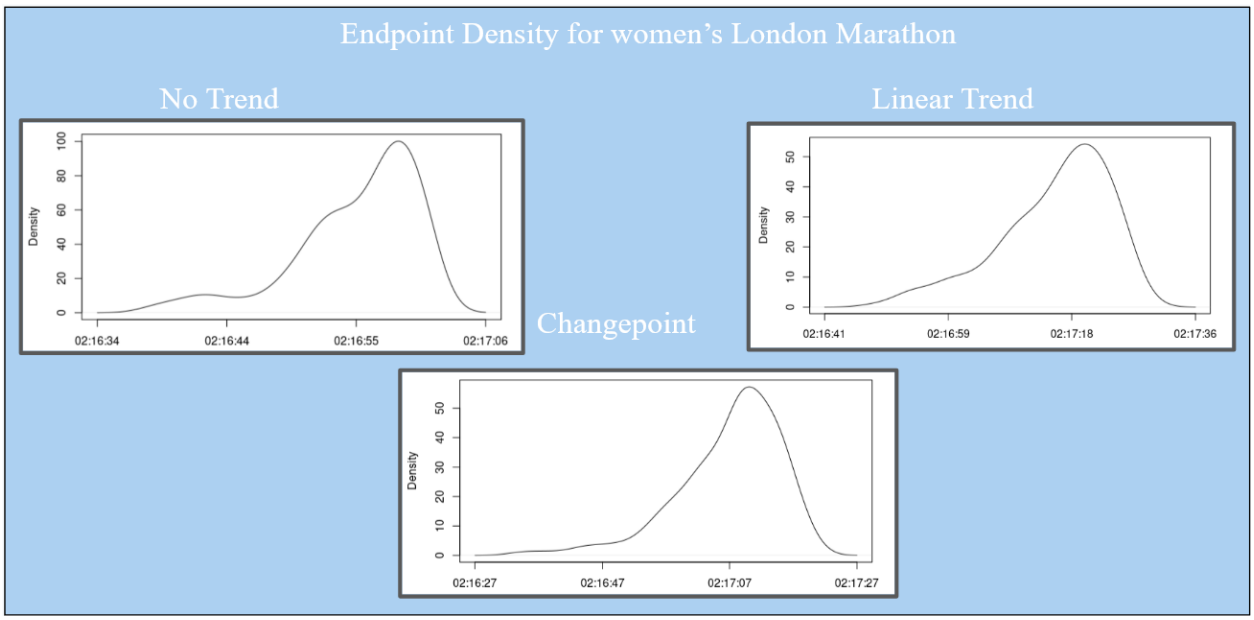


Figure 39: Density plots of the estimated endpoints, calculated from the output of the adaptive Metropolis algorithm for the women's London Marathon.

Considering the men's Boston Marathon, the no trend model estimates endpoints around 2 hours, 2 minutes and 41 seconds. The linear trend model estimates endpoints centered around 2 hours, 3 minutes, and 31 seconds, and the changepoint model estimates endpoints around 2 hours, 2 minutes, and 41 seconds. These results can be viewed in the visualizations below.

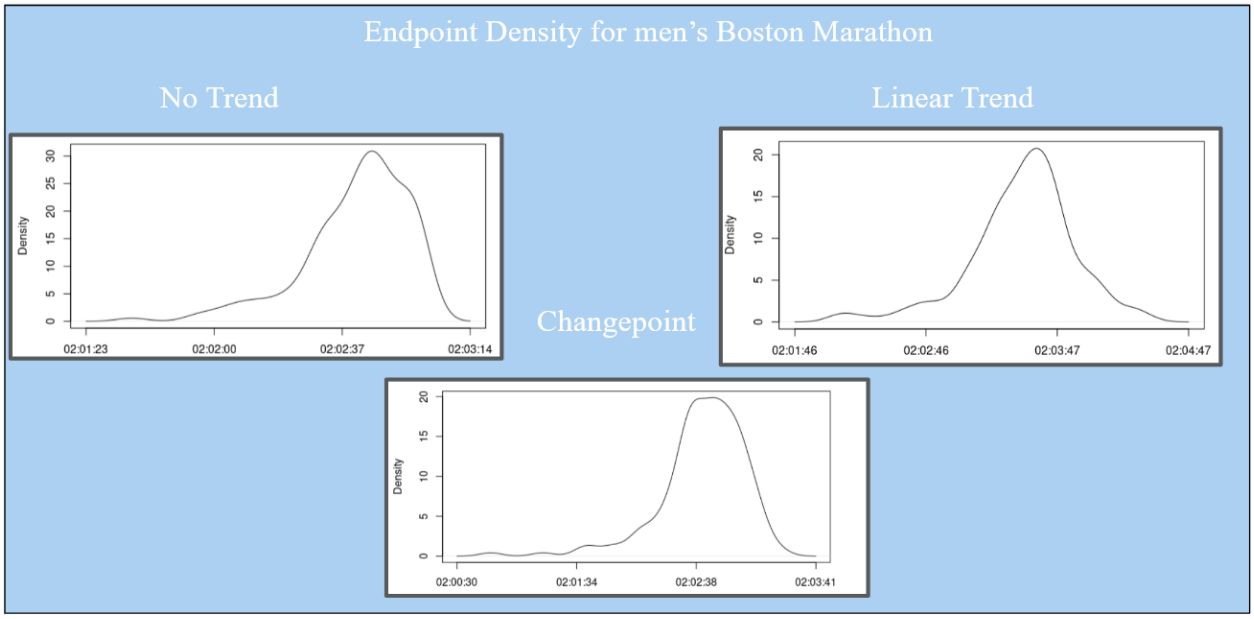


Figure 40: Density plots of the estimated endpoints, calculated from the output of the adaptive Metropolis algorithm for the men's Boston Marathon.

The women's Boston Marathon estimates endpoints around 2 hours, 18 minutes, and 45 seconds for the no trend model. Then, the linear trend model estimates endpoints around 2 hours, 19 minutes, and 47 seconds and the changepoint model approximates endpoints around 2 hours, 21 minutes, and 35 seconds. These calculations can be confirmed by the figures below.

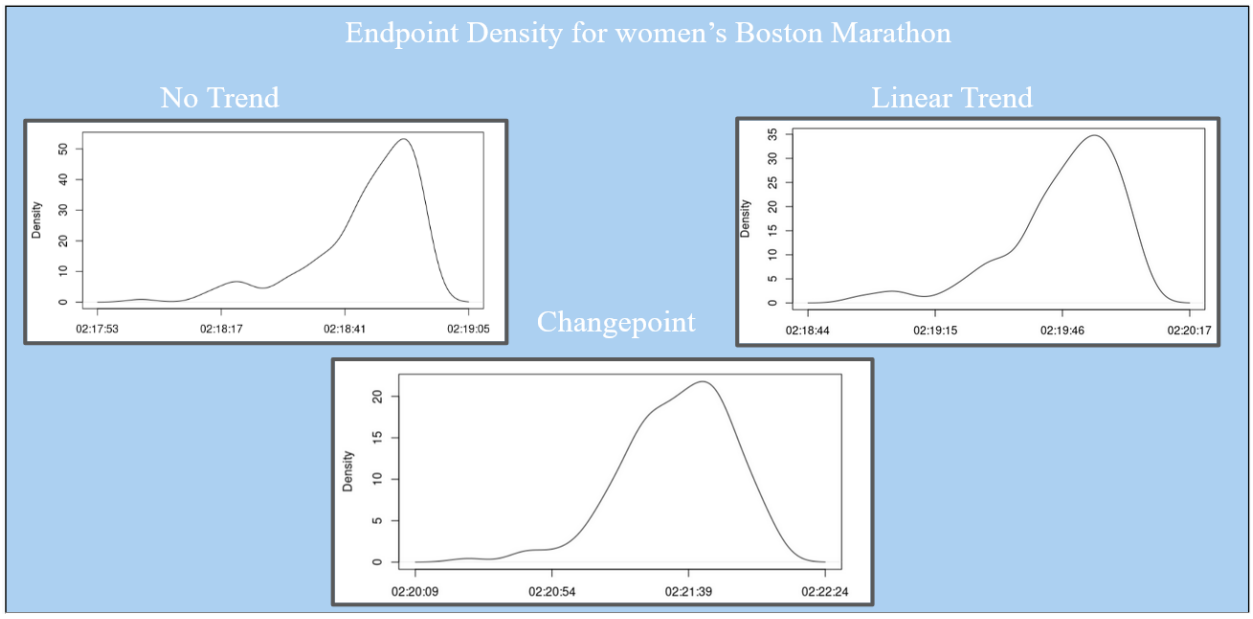


Figure 41: Density plots of the estimated endpoints, calculated from the output of the adaptive Metropolis algorithm for the women's Boston Marathon.

The men's Berlin Marathon estimates endpoints around 2 hours and 50 seconds for the no trend model, while the linear model predicts endpoints around 2 hours, 3 minutes, and 22 seconds. The changepoint model approximates endpoints around 2 hours, 3 minutes, and 04 seconds. These results can be validated by the figures below.

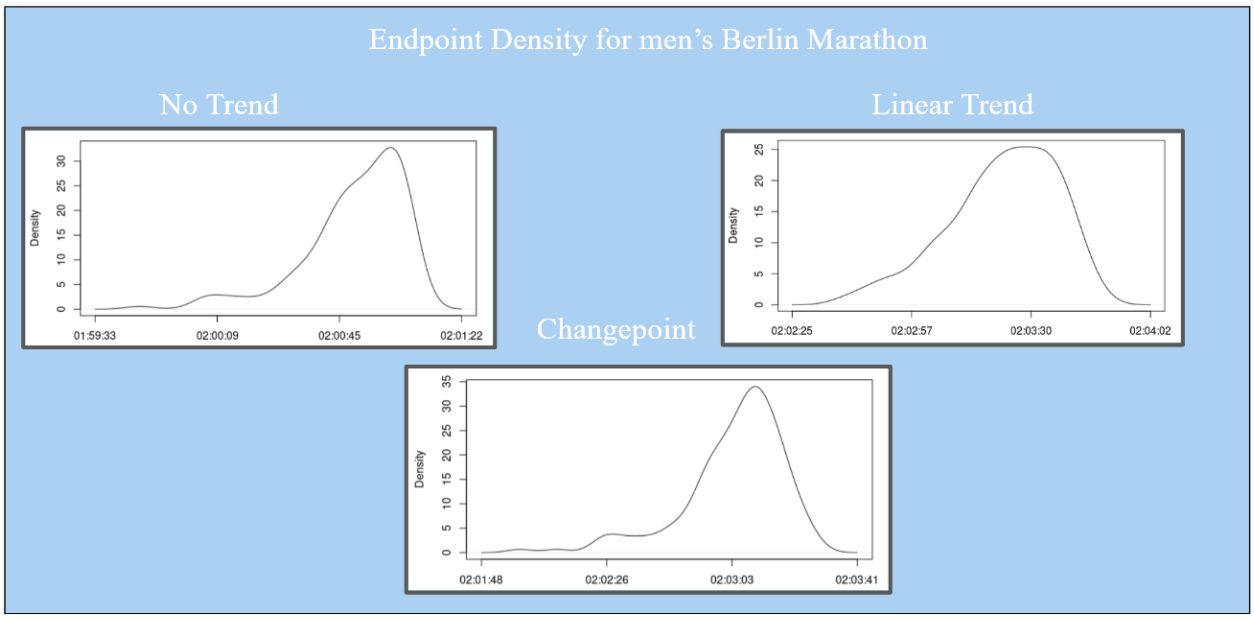


Figure 42: Density plots of the estimated endpoints, calculated from the output of the adaptive Metropolis algorithm for the men's Berlin Marathon.

Then, the no trend model for the women's Berlin Marathon estimates endpoints around 2 hours, 11 minutes, and 22 seconds. The linear trend model estimates endpoints around 2 hours, 16 minutes, and 20 seconds, where the changepoint model predicts endpoints to be around 2 hours, 16 minutes, and 35 seconds, as viewed below:

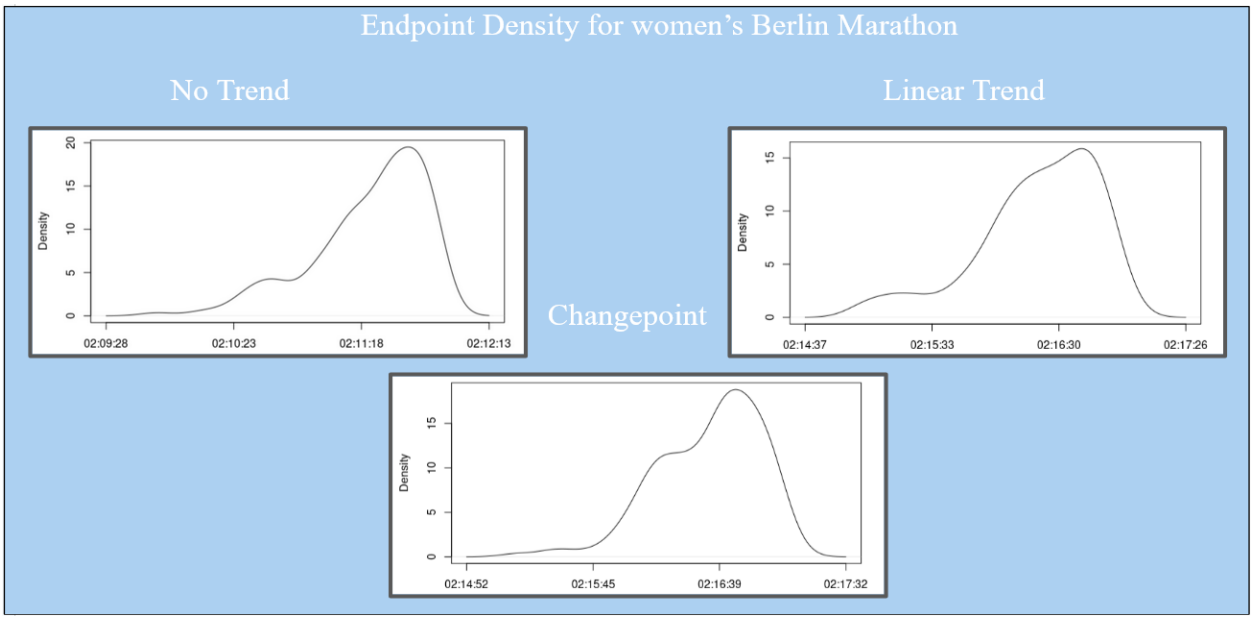


Figure 43: Density plots of the estimated endpoints, calculated from the output of the adaptive Metropolis algorithm for the women's Berlin Marathon.

It is also interesting to note the behavior of β_1 in the linear trend models, as well as the behavior of β_1 and β_2 in the changepoint models across races. These density plots are relatively similar across the various datasets, but can display valuable information about how race performances are limited. In most of the density plots of β_1 across the linear trend models, there is a sharp decrease at a point where β_1 values get more positive. This pattern can be represented by the plot below, displaying the density of β_1 for the men's Chicago Marathon linear trend model.

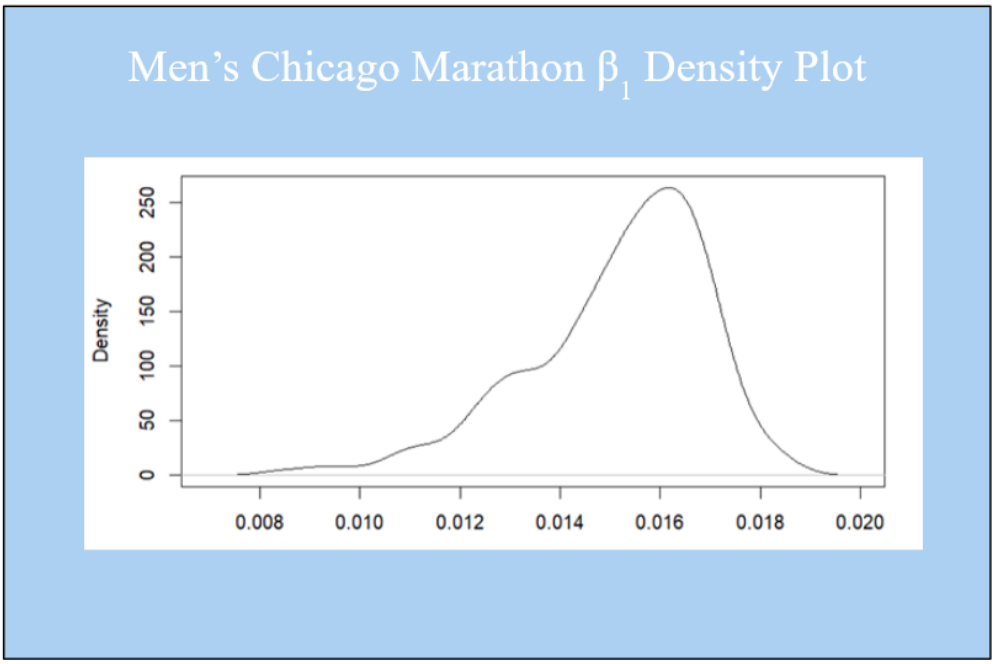


Figure 44: Density plot of β_1 from the adaptive Metropolis algorithm in the men's Chicago Marathon linear model.

Thus, this displays the limitation to how fast people can run according to the linear trend model, which is a practical result as runners are not infinitely getting faster. In addition, observing β_1 and β_2 trends across the changepoint models indicates a similar conclusion. As steep drop-offs are observed, such as in the below density plots of β_1 and β_2 in the women's Chicago Marathon changepoint model, it can be concluded there is in fact a limitation to how fast racers can be, even considering a changepoint in 2017, as is the claim here.

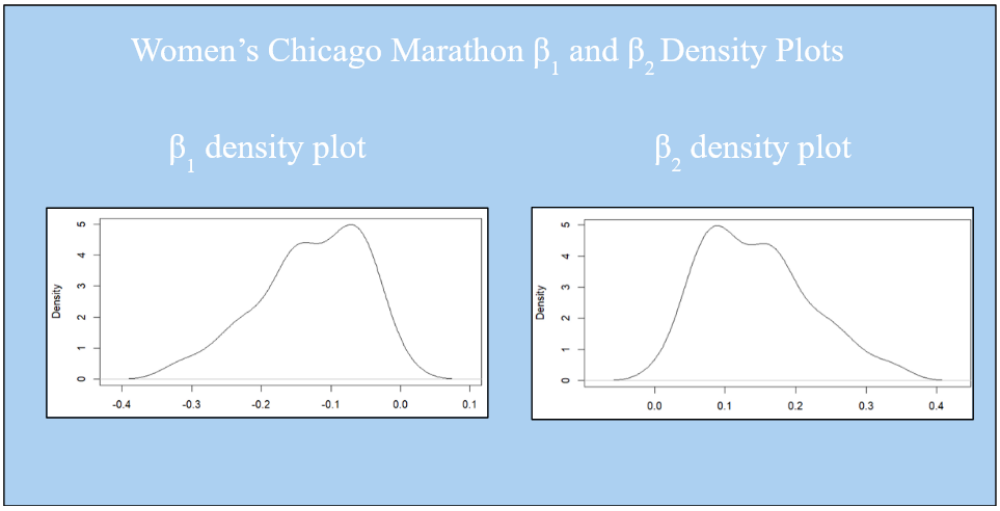


Figure 45: Density plots of β_1 and β_2 from the adaptive Metropolis algorithm in the women's Chicago Marathon changepoint model.

It should be noted that the above plots are similar across datasets, thus similar conclusions can hold for each race.

10. Discussion

Given the results above, we should note a few concerns. The first is a concern involving ξ. When $-1 < \xi < -.5$, maximum likelihood estimators are generally obtainable, but do not have the standard asymptotic properties [8, p. 55]. As noticed above, most of the ξ values fall within this range, although the rest of the results do not show any indication of a deviation from standard asymptotic properties.

In addition, these models have stability and sensitivity concerns, as the number of years greatly impacted the results of the analysis. However, given the variability of marathon races dependent on a variety of factors such as weather, participation, and competition, these issues were to be expected to some extent. For this particular analysis, we were able to mitigate sensitivity and stability concerns by gathering a sufficient amount of data. However, due to the reasons described surrounding race variability, these concerns will not completely disappear, regardless of the amount of data available.

Then, given ξ is negative in the above results, the endpoints calculated are upper end-points of the distribution [8, p. 55]. This supports the practicality of the sport, as running times have a limitation regardless of the race course, condition, or shoe technology available. Then, we can observe the endpoints found above and determine whether they are reasonable. In the following discussion, note the term "central endpoint" refers to the value corresponding to the mean of each posterior endpoint distribution.

Starting off with the men's Boston Marathon, the approximated central endpoints calculated by the

various models range from 2:02:41 to 2:03:31 (given in hours:minutes:seconds). The course record for the men's Boston Marathon is 2:03:02, set in 2011 by Geoffrey Mutai [7]. Thus, we notice the current course record falls just within the upper end-point of the distribution of running times estimated for 2024, given by all three models. This could indicate a high probability of estimated endpoints nearing the course record, and can serve as an expectation for the fastest 2024 performances. The winner of the 2024 Boston Marathon, Sisay Lemma, won the race in 2:06:17, noticeably outside of the range of estimated upper endpoints for the year. This indicates the 2024 men's Boston Marathon participants did not approach the upper end-point of times as estimated by any of the models. However, given the 2024 Boston Marathon experienced warmer temperatures than what can be observed in a typical year, we can suggest the winning time is outside the range of estimated upper end-points largely due to the weather.

A similar outcome can be observed for the women's Boston Marathon, with calculated central end-points ranging from 2:18:45 to 2:21:35. Given the women's Boston Marathon record is 2:19:59, set in 2014 by Buzunesh Deba [7], we notice the estimated endpoints for 2024 closely surround this value, particularly when considering the linear trend model. Thus, we can observe a high probability of estimated endpoints nearing the course record in 2024, indicating that the fastest results should reach, or even surpass, this time. However, the winner of the 2024 Boston Marathon was Hellen Obiri with a time of 2:22:37, indicating again that the racers underperformed, potentially due to warmer temperatures, according to the given estimates.

Observing now the men's London Marathon, we see the calculated central endpoints range from approximately 2:01:12 to 2:04:51. Kelvin Kiptum holds the current course record for the men with a time of 2:01:25, set in 2023. This is an indicator that the estimated endpoints for 2024 reach the cusp of the current record, but do not surpass it. Thus, we can assume the models estimate Kiptum's time to be the upper bound of the distribution of estimated endpoints in 2024, and breaking the record in 2024 is less probable than achieving slightly slower times. Indeed, the winner of the men's 2024 London Marathon was Alexander Munyao, with a time of 2:04:01. This is within the estimated range of endpoints, specifically considering the estimates derived from the linear trend model, thus sufficiently reflecting the estimated upper end-point of the distribution of running times in the 2024 men's London Marathon.

Concerning the women's London Marathon, we observe central endpoints ranging from 2:16:55 to 2:17:15. Prior to 2024, the women's course record was 2:15:25, set by Paula Radcliffe in 2003 [14]. We can notice this course record is significantly outside the range of estimated endpoints calculated by any of the models. This is likely due to the fact that the women's London Marathon analysis only includes data extending back to 2004, thus excluding the withstanding record set in 2003. In 2024, Peres Jepchirchir won the women's race with a time of 2:16:16, which is faster than what was estimated as the upper end-point of

the distribution of running times in 2024 for the women's London Marathon. This suggests the models lacked sufficient data to accurately estimate times, given they excluded a vital year which included the women's London Marathon course record.

The men's Berlin Marathon models resulted in estimated central endpoints ranging from 2:00:50 to 2:03:22. Eliud Kipchoge set a course record for the men's race in 2022 with a time of 2:01:09 [6]. This value falls within the estimated endpoints of the no trend model pertaining to the Berlin Marathon, although it is closer to the upper bound of these estimates than in prior races. Thus, in 2024, the estimated endpoints are expected to potentially near this record, but are highly improbable to surpass it, given by the slower estimates for the changepoint and linear trend models. In 2024, the Berlin Marathon winner, Milkesa Mengesha [5], won with a time of 2:03:17. This time falls within the estimated endpoints calculated for 2024 by the linear trend and changepoint models. Then, by these models, the winning performance in the 2024 men's Berlin Marathon successfully reached the estimated upper end-points of the distribution of running times.

In addition, it can be observed that the women's Berlin Marathon models resulted in estimated central endpoints ranging from 2:11:22 to 2:16:35. Thus, given the course record for the women's Berlin Marathon is 2:11:53, achieved by Tigist Assefa in 2023 [6], it is estimated that the endpoints for 2024 are expected to near the record, but are improbable to surpass it. Tigist Ketema won the 2024 Berlin Marathon with a time of 2:16:42 [5]. This result is just out of the range of the estimated central endpoints, but viewing the entire distribution of estimates, we notice this was still a probable result, particularly when viewing the linear trend and changepoint models.

To observe the men's Chicago Marathon, note the range of estimated central endpoints is given as 2:00:12 to 2:03:44. The men's course record was achieved in 2023 by Kelvin Kiptum with a time of 2:00:35 [13]. Once again, this value falls into the predicted range of the estimated endpoints for 2024, although it seems to be nearing the upper-limit of those estimations. The men's winner of the 2024 Chicago Marathon was John Korir with a time of 2:02:44. This falls well within the range given by both the changepoint and linear trend models, thus we can observe the winning performance in the men's 2024 Chicago marathon successfully reached the estimated upper end-points of that year's distribution of running times.

The women's Chicago Marathon models resulted in a range of predicted central endpoints from 2:12:57 to 2:17:43. Prior to 2024, the women's Chicago Marathon course record was achieved by Sifan Hassan in 2023 with a time of 2:13:44. Then, the estimated endpoints suggest times that with high probability would near, and possibly surpass, this record, particularly in the linear trend and no trend models. The current women's course record, and marathon world record, was achieved in 2024 by Ruth Chepng'etich with a time of 2:09:56, which is faster than the range of estimated endpoints by any of the models [13]. Although the

estimated endpoints suggested 2024 times that with high probability would near, and possibly surpass, the previous course record of 2:13:44, the three models did not estimate any times within the upper end-point of the distribution to be faster than 2:10:00. This indicates a limitation to estimating endpoints, which are not as flexible or accurate when considering their predictive abilities as other calculations.

Not only are endpoints important to estimate and validate specific performances, but they can also be utilized to predict future ones, especially with the presence of the prominent skew in each calculation. The endpoint densities can be used to calculate the probability of achieving a result, conditional on previous results or circumstances. This could be an area of future study, and something that may be interesting to peruse, especially as running times seem to be improving at an unusually rapid pace.

It should be noted the above endpoint calculations are somewhat surprising, given in many of the race analyses, the estimated endpoints for the linear trend and changepoint models are slower than the no trend models. However, given the significance of the linear trend models over the no trend models, as well as the positive β_1 values indicating faster times with a linear trend across analyses, the opposite should be true. The same can be said when comparing changepoint models to linear trend models across the various races. Thus, caution should be taken when interpreting the above endpoint results.

Besides the endpoint observations, we must also note the significance of each model in the results. With sufficient data, each race displayed significance of both the linear trend model (compared to the no trend model) and the changepoint model (compared to the linear trend model). Then, we can observe each race has evidence for achieving faster performances over time, given by the positive values of β_1 in the linear trend models. In addition, each race has even stronger evidence for a changepoint in 2017, corresponding to the official retail of the Nike Vaporfly 4%. To review from (Part One), a changepoint is a point in the data which corresponds to a significant change in the overall trend of the outcome. This can include properties such as changes in the slope of the outcome or a jump within the data, considering the same slope.

Thus, we have reasonable evidence to say the introduction of the Nike Vaporfly 4%, and succeeding super-shoes, could have had a reasonable impact on recent race performances in the Boston, Berlin, London and Chicago marathons. However, other factors could have played a role into this changepoint, given marathon races are variable and dependent on a multitude of factors described previously. Then, to not associate causality with super-shoes, we can instead suggest that their introduction correlates with recent improvements, beginning in 2017, specifically in the performances of the four marathons analyzed.

11. Conclusion

Not only are there obvious trends in marathon/half marathon data over time, as corresponding to

the annual records given in Part One, but we can also gain insight into what may be propelling recent major marathon performances by observing a larger amount of data from more specified races. One of the hypotheses we focused on in Part Two was the statistical impact of super shoes on major marathons in the last seven years.

Statistical significance was found for a changepoint in 2017, meaning although a general trend was also found in recent race performances of the Berlin, London, Boston, and Chicago marathons, there is even more evidence for something occurring in 2017 to change the trajectory of that trend. One possible explanation corresponds to the public introduction of super shoes in 2017 with the Nike Vaporfly 4%. The acceleration of improvement can be attributed to a multitude factors, many of which are hard to observe with low variability. Marathon races are extremely variable between years, depending on external factors such as weather, race condition, and competition. Thus, we can only suggest a possible explanation of the changepoint could be the introduction of super shoes, leading to improved performances in recent major marathons. Nevertheless, we cannot conclude this interaction is causal.

The results also led to an interesting discussion on the endpoints of each race, as well as the validity of current and possible future records. For future study, it would be interesting to further analyze endpoints for major marathons as technology continues to improve over time, including the calculation of various probabilities corresponding to setting a particular record.

The running community is dynamic, constantly catching up to current technologies, new training methods, and records. It is particularly interesting to see how these major marathons continue to respond to the acceleration of improvement in times, and how standards may change to reflect the speculated advantage of super shoes and additional technological advancements.

References

[1] Beat Knechtle et al. “World Single Age Records in Running from 5 km to Marathon”. In: *Frontiers* (2018). URL: https://doi.org/10.3389/fpsyg.2018.02013.

[2] Borja Muniz-Pardos et al. “Recent Improvements in Marathon Run Times are Likely Technological, Not Physiological”. In: *Sports Medicine* (2021). URL: https://doi.org/10.1007/s40279-020-01420-7.

[3] Jonathon W. Senefeld et al. “Technological Advances in Elite Marathon Performance”. In: *Journal of Applied Physiology* (2021). URL: https://doi.org/10.1152/japplphysiol.00002.2021.

[4] Hayden Bird. “Why a World Record Set in the Boston Marathon Wouldn’t Officially Count”. In: *The Boston Globe* (2023). URL: www.boston.com/sports/boston-marathon/2021/09/28/why-world-record-set-in-the-boston-marathon-wouldnt-officially-count/.

[5] BMW Berlin-Marathon. *50 BMW Berlin-Marathon 2024*. URL: https://berlin.r.mikatiming.com/2024/.

[6] BMW Berlin-Marathon. *BMW Berlin-Marathon Results and certificates*. URL: https://www.bmw-berlin-marathon.com/en/your-race/results.

[7] Boston Athletic Association. *Boston Marathon Archives*. URL: http://registration.baa.org/cfm_Archive/iframe_ArchiveSearch.cfm.

[8] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. 1st ed. Springer Series in Statistics. Springer London, 2001. ISBN: 978-1-85233-459-8.

[9] Trace A. Dominy and Dustin P. Joubert. “Effects of a Carbon-Plated Racing Shoe on Running Economy at Slower Running Speeds”. In: *International Journal of Exercise Science: Conference Proceedings* 2.14 (2022). URL: https://digitalcommons.wku.edu/ijesab/vol2/iss14/15.

[10] Eric Gilleland and Richard W. Katz. “extRemes 2.0: An Extreme Value Analysis Package in R”. In: *Journal of Statistical Software* 72.8 (2016), pp. 1–39. DOI: 10.18637/jss.v072.i08.

[11] Eero Saksman Heikki Haario and Johanna Tamminen. *An Adaptive Metropolis Algorithm*. Vol. 7. No. 2. International Statistical Institute and the Bernoulli Society for Mathematical Statistics and Probability, 2001, pp. 223–242.

[12] Rob Hodgetts. “Eliud Kipchoge’s record-breaking Nike shoes to be banned”. In: (2020). URL: https://www.cnn.com/2020/01/31/sport/nike-world-athletics-vaporfly-alphafly-ban-spt-intl/index.html.

[13] Marathon Guide. *Chicago Marathon - Race Results*. URL: https://www.marathonguide.com/results/browse.cfm?MIDD=67241013.

[14] Marathon Guide. *London Marathon - Race Results*. URL: https://www.marathonguide.com/results/browse.cfm?MIDD=16240421.

[15] Emily Shapiro. "Boston Marathon runners brave brutal cold, wind and rain". In: (2018). URL: https://abcnews.go.com/US/boston-marathon-runners-brave-brutal-cold-wind-rain/story?id=54498734.

[16] Lars-Christian Simonsen. "The History of Nike Vaporfly: Changing the Game in Distance Running". In: (2023). URL: https://run161.com/shoes-gear/the-history-of-nike-vaporfly-changing-the-game-in-distance-running/.

[17] Richard Smith. "A Survey of Nonregular Problems". In: *Proceedings of the 47th session of the International Statistical Institute*. Paper C1-27. Paris, 1989, pp. 353–372. URL: https://doi.org/10.1002/sim.4780100603.

[18] "Super Shoes Explained". In: *Gazelle Sports* (2023). URL: https://gazellesports.com/blogs/news/super-shoes-explained?srsltid=AfmBOoriQYKWvVJk9KwMrNOQHayXemfWAbrk-1969EKHPfb76V06zXRC.

[19] World Athletics. *Half Marathon*. 2021. URL: https://worldathletics.org/disciplines/road-running/half-marathon.

[20] World Athletics. *New athletic shoe regulations approved by Council*. 2021. URL: https://worldathletics.org/news/press-releases/new-athletic-shoe-regulations-approved-2022.

[21] World Athletics. *All time Top lists*. 2024. URL: https://worldathletics.org/records/all-time-toplists.

[22] G. A. Young and Richard L. Smith. *Essentials of Statistical Inference*. Vol. 16. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2005. ISBN: 978-0-521-54866-3.