Quantifying Uncertainty in Marathon Finish Time Predictions

Anonymous Author(s)
Affiliation
Address
email

Abstract

In the middle of a marathon, a runner's expected finish time is commonly estimated by extrapolating the average pace covered so far, assuming it to be constant for the rest of the race. These predictions have two key issues: the estimates do not consider the in-race context that can determine if a runner is likely to finish faster or slower than expected, and the prediction is a single point estimate with no information about uncertainty. We implement two approaches to address these issues: Bayesian linear regression and quantile regression. Both methods incorporate information from all splits in the race and allow us to quantify uncertainty around the predicted finish times. We utilized 15 years of Boston Marathon data (312,805 runners total) to evaluate and compare both approaches. Finally, we developed an app for runners to visualize their estimated finish distribution in real time.

1 Introduction

A marathon is a long-distance road race where runners each complete 26.2 miles (42.195km). Many marathons, especially larger ones, can have tens of thousands of runners racing at once, with a significantly large number of spectators watching the race and cheering on the sidelines. Often, a task informally performed by those watching (or even those running the race) is predicting what time a given runner will finish the race. As spectators usually remain at one spot along the 26.2 mile course, they usually are only able to see a runner once. Thus, there is very limited information to make finish time predictions for the multiple hour race. Many marathons, however, use a chip in each runner's bib to track when runners complete certain portions of the race, often at every 5km increment. These in-race splits are often reported with the runner's finish time, and are occasionally even posted live as the race is happening.

Using a runner's splits gives spectators a path to make live predictions for their finish time. Traditionally, when major marathons display estimated finish times, the common approach is to utilize only the average pace shown from the most recently taken split. In this prediction, the pace is assumed to be held constant for the rest of the race and is extrapolated to arrive at a prediction. Predictions like this can be helpful for getting a general sense of when a runner will finish, but they have two key issues.

First, the estimates do not consider the in race context that can determine if a runner is likely to finish faster or slower. For example, marathon runners are commonly known to run slower during the second half of a race due to accumulated fatigue, and thus the traditional prediction method will underestimate the finish time. Second, the prediction is a single point estimate that has no additional information about the uncertainty behind the estimate. Intuitively, we should feel more confident about a prediction made when a runner has completed 30km of the race (about 75%) rather than a prediction made when the runner has only completed 10km (about 25%). Our predictions can reflect the uncertainty behind a point estimate with a range of possible finish times, which should be narrower and more precisely around an estimate as the runner gets closer to the finish of the race.

Preprint. Under review.

Seeking to address these two issues, we identified two approaches: Bayesian linear regression and quantile linear regression. Both methods incorporate multiple pieces of information from the race and allow us to quantify uncertainty around the predicted finish times.

2 Data

We focused our analysis on the Boston Marathon, the worst oldest annual marathon and a World Marathon Major. The event hosts tens of thousands of runners every year. Most runners qualify to compete in the marathon by hitting notoriously difficult standards, while the rest of the field is made up of charity runner spots, which have no qualification standards. We scraped data from the website of the Boston Athletic Association (BAA), the organization that hosts the marathon [4]. Our dataset contains the name, age, gender, and in-race splits (5K, 10K, 15K, 20K, HALF, 25K, 30K, 35K, 40K, and FINISH, all in seconds) for every finishing runner of the Boston Marathon from 2009-2023 (n=312,805 total). We partitioned this data into a training set (286,777 runners from 2009-2022) and a test set (26,028 runners from 2023). The distribution of finish times is shown in Figure 1.

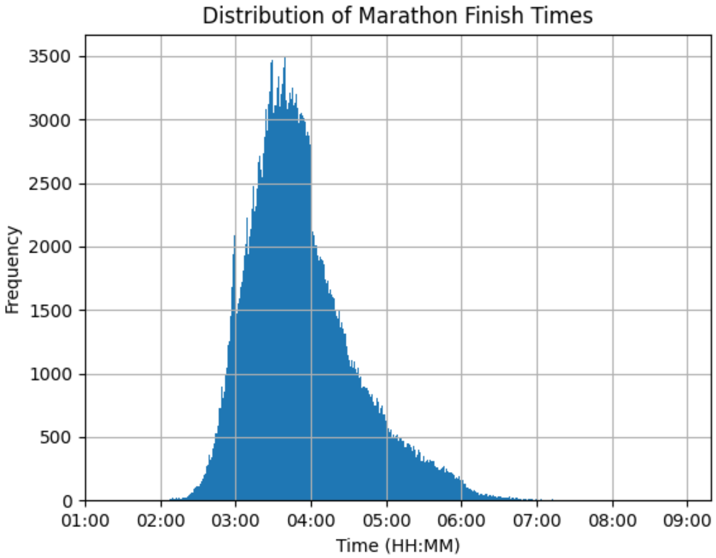


Figure 1: Finish time distribution of the training set (2009-2023)

By reformatting the data, we can better represent the above features for our prediction task. For each stage of the race, we can compute the *total_pace* for each individual. *Total_pace* is the average pace covered by the runner up until that stage of the race, and it forms the basis of the traditional method, which assumes that pace will be held constant for the rest of the race. In Figure 2, we directly compare true finish times with extrapolated *total_paces*, which represents the traditional method's finish time estimates. The red line represents the condition where the traditional method accurately predicts the finish time. For each of the different stages, most of the points lie above the traditional estimate line. We also plotted the best fit lines for each of the stages, and each one visually reflects a different relationship than the traditional method. In the next sections, we explore different possible relationships that could be used to predict finish times, and evaluate the performance of each of the models.

3 Methods

The traditional method of extrapolating the current pace (**Method A**) is used as a baseline. We considered three additional possible linear relationships between a runner's splits and their finish time:

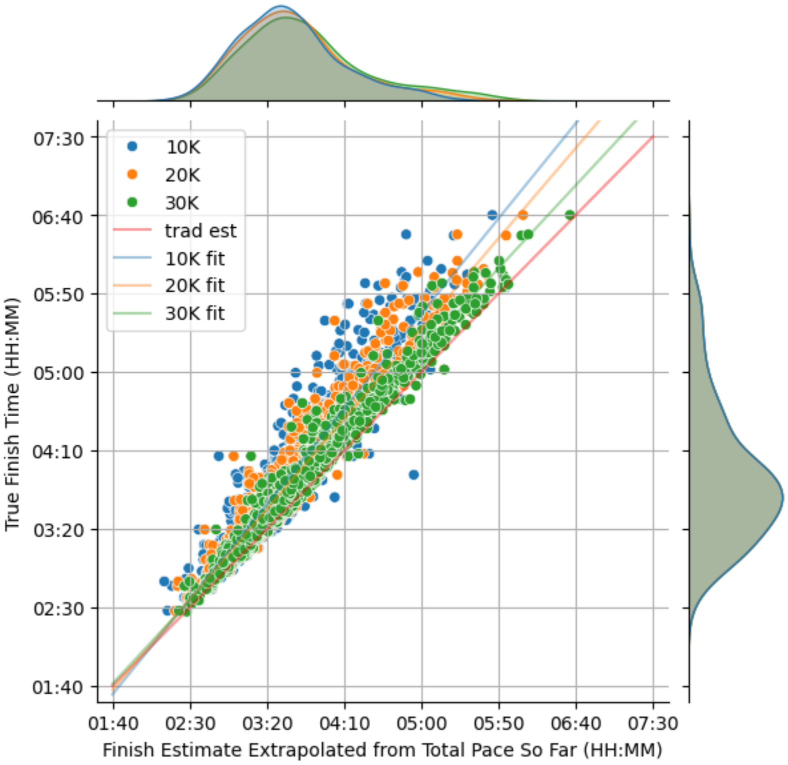


Figure 2: For three stages of the race (10K, in blue, 20K, in orange, and 30K, in green), the finish estimates extrapolated from the total pace so far (x-axis) are compared to the actual finish times (y-axis). The red line depicts the traditional estimate, while the other lines are the best fit lines for each of the three stages of the race.

Method B: the overall pace so far (*total_pace*) is used as a single predictor
Method C: the overall pace so far (*total_pace*) and the "current pace" (the pace of the most recent 5K) are the two predictors
Method D: all previous splits in the race are used as predictors

In addition, we explored two different models. The Bayesian linear regression [1] model incorporates Bayesian statistics by placing priors on the parameters of the linear regression model. By combining a likelihood function with a specified prior distribution, we form a posterior distribution of possible finish times for a given individual. We interpret the posterior both as a point estimate (using the median of the distribution) and a credible interval: a central region of the distribution we can use to quantify uncertainty. The Bayesian linear regression model is built using PyMC, a Python library for creating Bayesian models [6]. Quantile linear regression [5] takes a different approach. In contrast to regular linear regression, which estimates the mean of the response variable, quantile regression estimates quantiles. Thus, we use the estimated median as a point estimate for finish times, and create credible intervals using the corresponding quantiles.

We can evaluate the performance of each combination of method (**A** through **D**) and model (Bayesian linear regression or quantile regression) by training the model on the training set (2009-2022) and predicting and computing the root mean squared error (RMSE) of the runners in the test set (2023).

Increasing the amount of information (number of features) used to predict finish times should lower the RMSE and improve the accuracy of the predictions. However, we note that Method D, the method with the most prior information, has strong issues with collinearity, as a runner's previous splits are strongly correlated with each other.

We found that adding more features to the Bayesian model led to significantly longer model runtimes. Only Methods B and C had reasonable runtimes for the Bayesian model, so we decided to implement

Methods B and C for both Bayesian and quantile regression and compare the performances to the baseline. In addition, we subsampled our training set to randomly select 5000 runners from 2022, to speed up runtime while still having a reasonably sized dataset to make inference with. Labels in the below plots are as follows:

extrap: the baseline, traditional extrapolation prediction (Method A)
bayes1: Bayesian linear regression using just the overall pace as a predictor (Method B)
bayes2: Bayesian linear regression using both the overall pace and the "current pace" as predictors (Method C)
quant1: Quantile linear regression using just the overall pace as a predictor (Method B)
quant2: Quantile linear regression using both the overall pace and the "current pace" as predictors (Method C)

4 Results

As shown in Figure 3, all four of our models (bayes1, bayes2, quant1, and quant2) have very similar test RMSE at most levels of the race. With the exception of bayes2 and quant2 at the 40K mark, all models improve upon the traditional method at all levels, and significantly outperform it in the beginning and middle stages of the race. It is especially important to have better finish estimates earlier in the race, as there is the greatest amount of uncertainty at these stages. The gap in RMSE decreases between the traditional model and the bayes1 and quant1 models as the race gets closer to finishing, and all 3 seem to have approximately the same RMSE at the 40K mark. This makes sense because the traditional model also benefits from decreased uncertainty at the latter stages of the race, as the overall pace so far becomes a better estimator of the true overall finish pace we want to indirectly estimate.

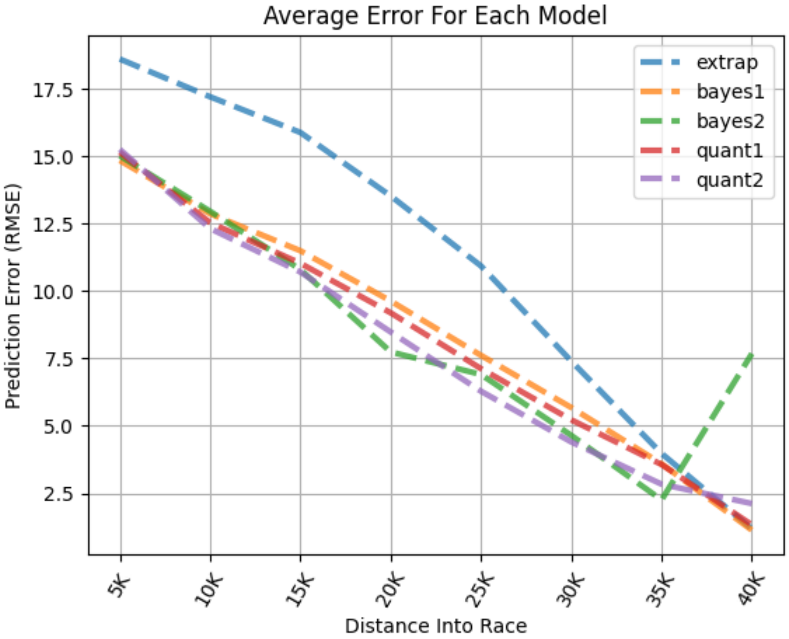


Figure 3: RMSE at different stages of the race for the traditional method (blue), both Bayesian linear regression models (orange and green) and both quantile linear regression models (red and purple).

The Method C models (bayes2, quant2) seem to do slightly better at most stages compared to the Model B models (bayes1, quant1), although that gap is small compared to the gap between all four of those models and the traditional one. Intuitively, this makes sense because the additional predictor is enough information to get a slightly better estimate. The bayes2 and quant2 models perform relatively poorly at 40K when compared to performance at the other stages. A possible explanation for this could be that the total pace could be so much more important of a predictor than the current pace at this stage of the race.

The benefits of our models lie not only with improved average accuracy of the finish time point estimates. For each individual, these models also provide a credible interval, used to quantify uncertainty behind the estimate. When passing in a student's feature predictors at a given distance into one of the models, we can create an X% credible interval $[t_1, t_2]$ such that that the true finish time falls between t_1 and t_2 X% of the time.

In order to validate these credible intervals generated from the models, we perform checks to see how well they fit our assumptions. Specifically, we examine the credible interval sizes. On average, we expect the credible interval sizes to decrease as one gets further into the race, which fits with our intuition that one should be more certain of the estimate as they get closer to finishing. Figure 4 shows that this is generally true for three different credible intervals (50%, 80%, and 95% intervals), as each average interval size decreases and converges towards 0 as the race progresses and gets closer to finishing. The only exception is with the bayes2 model at the very beginning (5K mark) and end (40K mark) of races.

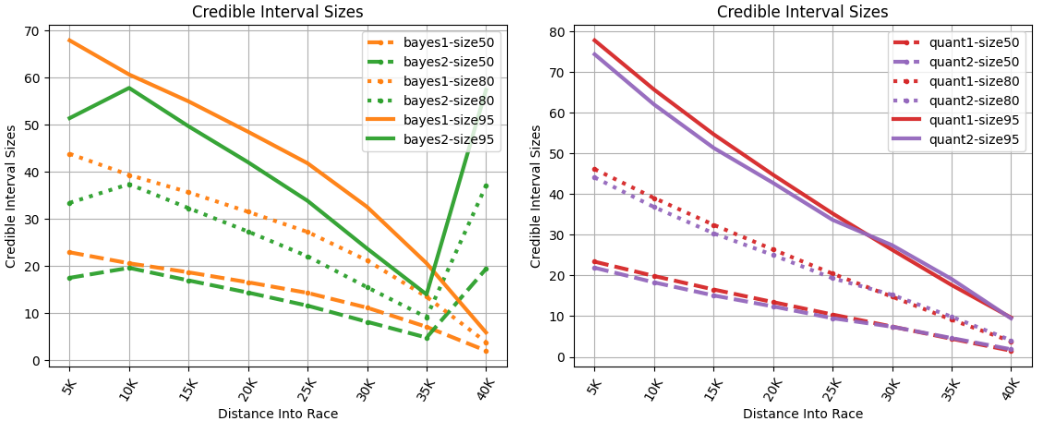


Figure 4: Average credible interval sizes at each stage of the race for models bayes1 (left, orange), bayes2 (left, green), quant1 (right, red) and quant2 (right, purple)

We also want to see if, for a given X% interval, approximately X% of runners actually finish within that interval. This gives us an approximation to our true goal that an individual's finish time has an X% chance of being within that predicted interval. Figure 5 shows the proportions of intervals that contain the true value across different stages of the race for each model. We see that the proportions are roughly around the expected proportions of 50%, 80%, and 95%, but do not match up perfectly. Inspecting the graph, we see that over time, for a given interval size, the bayes1 and bayes2 models seem to increase above the expectation over time. In contrast, for the quant1 and quant2 models, the proportions peak and are above the expectation in the middle of the race (15K through 30K), and are lower than the expectation at the beginning and end of the races.

5 Application

We developed an application to display how the bayes1 model can be used to make predictions for a marathon race in real time. The *My Plot* tab of the app can be used to "simulate" a race; a user can sequentially enter in splits (in increments of 5K) and the app will dynamically compute and display finish time statistics. One output is a plot, displaying the predicted finish time probability distributions at different stages of the race. The centers of each curve represent the most probable finish times at that point of the race, and seeing multiple distributions together visually shows how the prediction changes over time. A narrower distribution represents more precise predictions and narrower credible intervals. The other output is a table showing the median finish time prediction as well as credible intervals (50%, 80%, and 95%) for each stage of the race. This view of the data allows for a more detailed view of the actual values from the prediction. The other tab (*NUCR Plots*) shows the tables and plots of a few runners that ran the 2023 Boston Marathon, using the splits from their races.

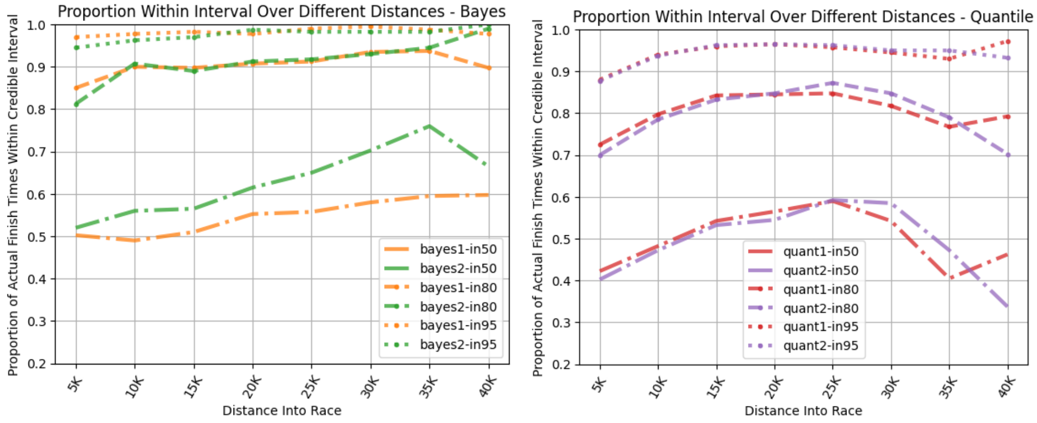
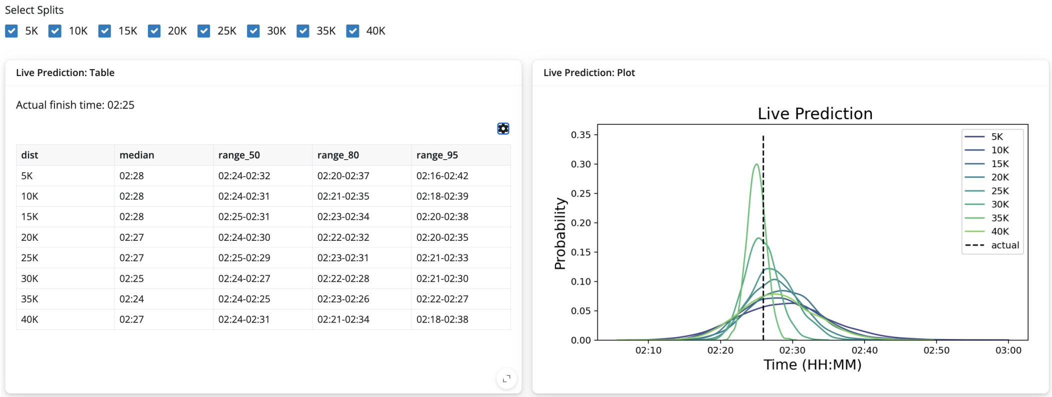


Figure 5: Proportion of true finish times falling within credible intervals at each stage of the race for models bayes1 (left, orange), bayes2 (left, green), quant1 (right, red) and quant2 (right, purple)



Actual finish time: 02:25

dist	median	range_50	range_80	range_95
5K	02:28	02:24-02:32	02:20-02:37	02:16-02:42
10K	02:28	02:24-02:31	02:21-02:35	02:18-02:39
15K	02:28	02:25-02:31	02:23-02:34	02:20-02:38
20K	02:27	02:24-02:30	02:22-02:32	02:20-02:35
25K	02:27	02:25-02:29	02:23-02:31	02:21-02:33
30K	02:25	02:24-02:27	02:22-02:28	02:21-02:30
35K	02:24	02:24-02:25	02:23-02:26	02:22-02:27
40K	02:27	02:24-02:31	02:21-02:34	02:18-02:38

Figure 6: Screenshots from the application. On the left, a table showing the median prediction as well as credible intervals (50%, 80%, and 95%) at each stage of the race. On the right, the probability distributions for finish times at each stage of the race shown in a plot.

The *My Plot* tab can be a helpful tool to understand the distribution of possible finish times as the race is occurring. This can be used in a variety of contexts, whether for a coach informing a runner mid-race that they are on-pace for their goal or not, a spectator trying to assess when their friend or family member will cross the finish line, or even the runner themselves after the race analyzing how well they adhered to their race strategy.

The app can be found at

`https://bonyejekwe.shinyapps.io/marathon_predictor/`

6 Conclusion and Future Work

Both Bayesian linear regression and quantile linear regression can be used to address the issues present from the traditional method of estimating marathon finish times. Both benefit from significantly improved point estimates by taking into account the context of in-race splits, while simultaneously providing additional context around the estimate with credible intervals to provide a sense of uncertainty. What remains to be seen, however, is if there are better choices for predictors that would provide the models with even better estimates. While adding features significantly increases the time it takes to fit the Bayesian models, adding well-chosen features should help get better estimates. We can avoid the collinearity issues we discussed with Method D above by utilizing other features we have available to use in our dataset, such as age and gender. In future work, we could better quantify the effects of feature selection on the overall RMSE for each model.

The application currently only implements the bayes1 model for simplicity. We decided that implementing multiple models could be overwhelming for the user, distracting from our goal to get fast and accurate information. However, by incorporating all models, there would be a visual comparison between models, and could possibly reveal nuanced differences in the models if they give vastly different results for the same user.

Finally, the model and resulting application described in this paper were developed for a specific use case: predicting marathon finish times using a model trained on Boston Marathon data. Changing the dataset to the results of a different marathon will alter the predictions to make the model more applicable towards that specific race. A future goal would be to add different major marathons to the application to allow runners to specify which race they are running, and update the model accordingly. The prediction task can even be adapted towards different goals. For example, the model can be modified to predict when a runner will cross a certain point in the race (say, the 30km mark) instead of the finish, which can be helpful for a spectator stationed at that point wanting to know when a specific runner will pass by.

All of the code used to create the analyses and develop the application can be found here: `https://github.com/bonyejekwe/Marathon_Predictor`

References

[1] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* Chapman and Hall/CRC.

[2] Allen, E. J., Dechow, P. M., Pope, D. G., & Wu, G. (2017). Reference-dependent preferences: Evidence from marathon runners. *Management Science, 63*(6), 1657-1672.

[3] Kwong, H. S., & Nadarajah, S. (2019). Modelling dynamics of marathons–A mixture model approach. *Physica A: Statistical Mechanics and its Applications*, 534, 120798.

[4] *Results: Boston Athletic Association.* Results | Boston Athletic Association. (2024). `https://www.baa.org/races/boston-marathon/results`

[5] Perktold, J., Seabold, S., & Taylor, J. (2024). *Quantile regression.* Quantile regression - statsmodels 0.15.0 (+431). `https://www.statsmodels.org/dev/examples/notebooks/generated/quantile_regression.html`

[6] *GLM: Linear regression.* GLM: Linear regression - PyMC 5.16.2 documentation. (2024). `https://www.pymc.io/projects/docs/en/stable/learn/core_notebooks/GLM_linear.html`