Marathon Race Time Prediction

Kexin Shi

2024-12-07

Table of contents

Summary

Over the last ten years, marathon running has become a popular physical activity around the world. This study aims to investigate the relationship between marathon runners' maximum distance ran per week during race training and their race time. We modeled a simple linear regression to predict marathon times based on their training patterns and tested on the model with the test dataset. By having this analysis, it helps marathon lovers to gain insights into how training volume influences race performance and to better prepare the race.

Introduction

Over the last ten years, marathon running has become a popular physical activity around the world Zoladz and Nieckarz (2021). It is commonly known as a high-level endurance exercise that requires the runners to have dedicated training Kaufmann et al. (2020). The maximum distance ran per week during race training is a key metric that marathon lovers would care about and is a key reference to predict which athletes will perform better than others. Thus, this study will investigate how the maximum distance ran per week during race training will predict the time it takes a runner to finish the race. Specifically, we want to answer the question: What predicts which athletes will perform better than others? How the maximum distance ran per week (in miles) during race training predicts the time it takes a runner to finish the race? This study uses the dataset from a a public dataset on GitHub, containing 13 variables about runners, such as age, bmi, maximum training distance per week (max, in miles) and actual marathon race time (time_hrs, in hours) and so on.

```
## Data Validation Check
### Correct Data File Format
#| include: false
#| echo: false
if (!inherits(marathon, "data.frame")) {
  stop("The data file is not in the correct format. Expected a CSV to be read as a data frame
}

empty_rows <- apply(marathon, 1, function(row) all(is.na(row)))
if (any(empty_rows)) {
  warning("There are completely empty observations. Consider removing these rows.")
  marathon <- marathon[!empty_rows, ]
}

if (any(duplicated(marathon))) {
  warning("There are duplicate observations. Removing them now.")
  marathon <- marathon[!duplicated(marathon), ]
}
```

EDA

We want to predict race time (in hours) (time_hrs) given a particular value of maximum distance ran per week (in miles) during race training (max). With this subset, we can plot a scatterplot to assess the relationship between these two variables.

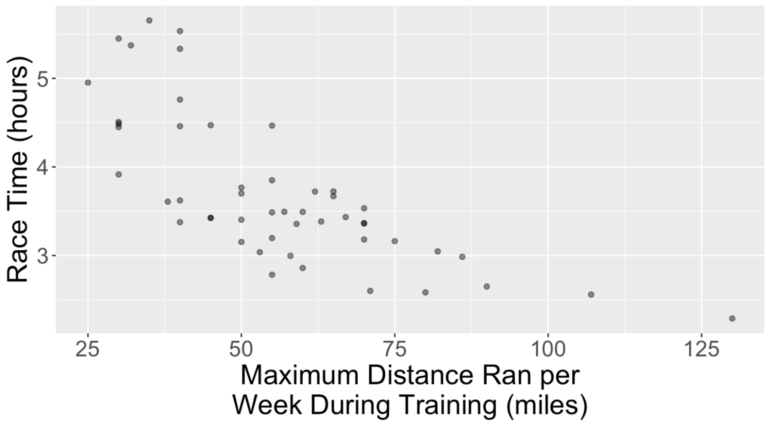

Figure 1: Scatterplot of Sub Dataset Maximum Distance Ran per Week vs. Race Time

Here's the full dataset of values.

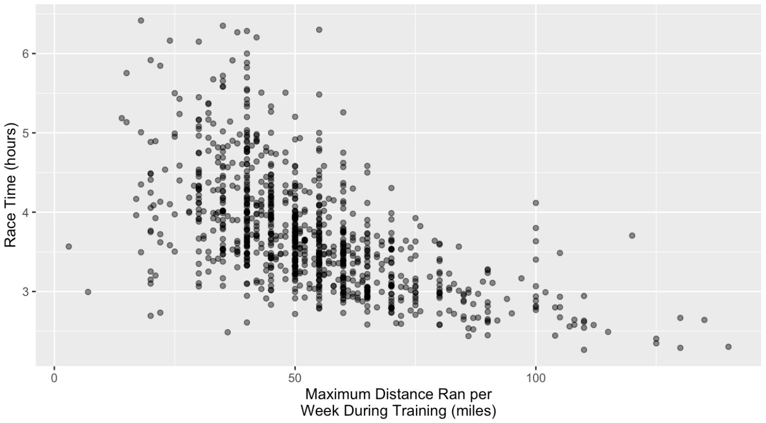

Figure 2: Scatterplot of Full Dataset Maximum Distance Ran per Week vs. Race Time

Analysis

We will analyze the data using simple linear regression to assess the relationship between maximum weekly distance and marathon race time. We will first split the dataset into the training and testing datasets, using 75% of the original data as the training data. The training set was used to fit the model, while the test set was used for performance evaluation. In the strata argument of the initial_split function, we will use the variable we are trying to predict.

Training Data

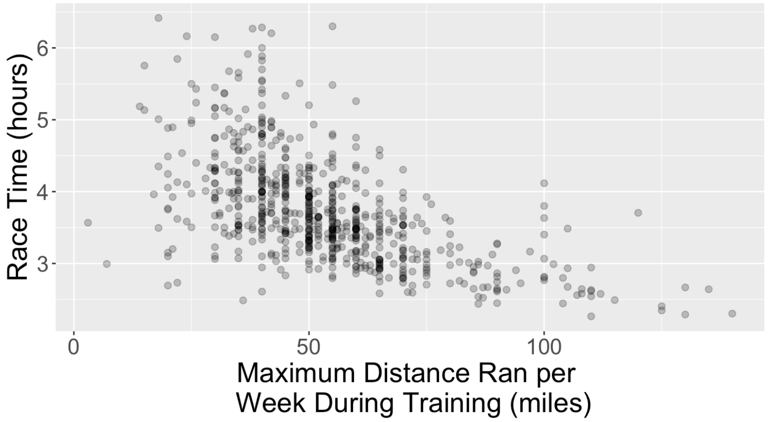

Figure 3: Scatterplot of Training Dataset Maximum Distance Ran per Week vs. Race Time

We can look at Figure 3 to assess the relationship between race time (time_hrs) and maximum distance ran per week during training (max) using only the observations in the training dataset.

Linear Regession

Now that we have our training data, the next step is to build a linear regression model specification.

After we have created our linear regression model specification, the next step is to create a recipe, establish a workflow analysis and fit our simple linear regression model.

Results

Now, let's visualize the model predictions as a straight line overlaid on the training data.

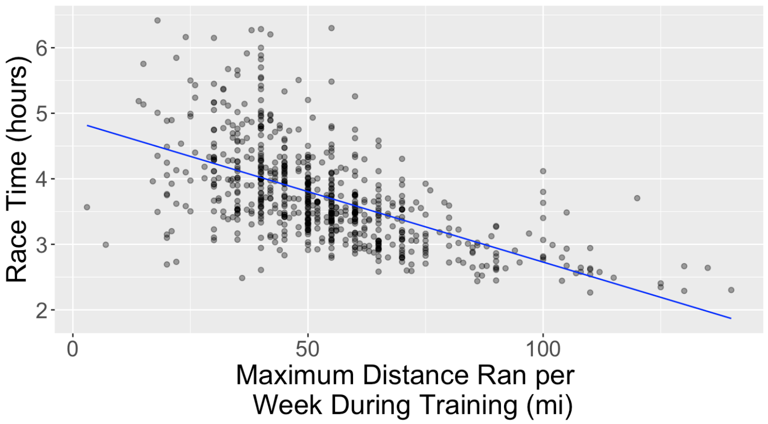

Figure 4: The Linear Regression of Maximum Distance Ran per Week And Race Time On Training Set

Model Performance

We can look at our model performance by looking at the RMSPE on the test data.

Table 1: Linear Regression ON RMSPE Metric

x
0.5504829

The Table 1 stands for the Root Mean Squared Prediction Error, which is a scoring metric that reflects how right/wrong each prediction is. It measures the distance of the prediction from the actual values on the test data. Lower RMSPE values indicate better predictive performance, meaning that the model's predicted values are close to the actual observed values Taraji et al. (2017). In our case, the RMSPE represents the average error in predicting marathon race time (in hours) for the runners. From the result of 0.5504829, This means that, our simple linear regression model's predictions deviate from the actual race times by approximately 0.55 hours on average.

Prediction on test data

Visualize the model predictions as a straight line overlaid on the test data

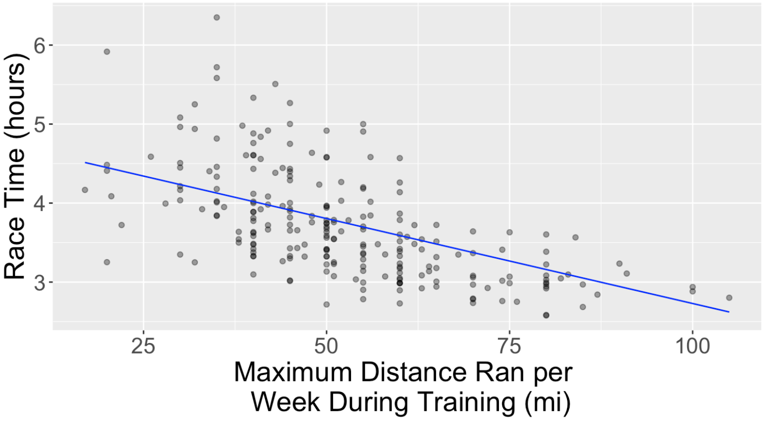

Figure 5: The Linear Regression Model Prediction Over Test Data

Table 2: Linear Regression Results

term	estimate	std.error	statistic	p.value
(Intercept)	4.8794113	0.0651428	74.90328	0
max	-0.0215038	0.0011449	-18.78199	0

Interpretation

By having Table 2, the slope coefficient is -0.0215, which shows a negative relationship between the maximum distance ran per week and the race time. As the maximum distance ran increases by every additional mile per week, the race time will decrease by 0.0215 hours. It shows that adding more training distance per week is beneficial for improving marathon performance. The intercept is 4.88, wich represents that when the maximum distance ran per week (max) is zero miles, the predicted race time is 4.88 hours. Although this scenario is unrealistic, it serves as the baseline value from which changes in race time are predicted based on the training distance. These coefficients with p-value of 6.03e-64 and 0 (less than 0.05) are statistically significant, meaning there is strong evidence that increasing the training mileage leads to faster marathon times.

If we want to manually calculate the marathon time, the formula will be like $times_{hrs} = 4.88 - 0.0215 \times max$, where $times_{hrs}$ is the race time and the max is the maximum distance (in miles) ran per week during training.

Discussion

However, there are several limitations of the model. As we are trying to find what predicts which athletes will perform better than others, other features about the runners may also be crucial to play roles in affecting the race time, such as their age, bmi. Therefore, we may need to add more features to the model to better predict the race time. Also, these features may not in a linear relationship with the race time, so using a Random Forest model to handle the complex, non-linear interactions between features may be useful.

To know whether the new model would be better than the simple linear regression or not, we can use those scoring metrics such as Mean squared error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and R-squared (R^2) on a test set or through cross-validation.A lower MSE, RMSE, MAE, MAPE would indicate that the new model makes more accurate predictions compared to the simple linear regression model. A higher R-squared would suggest that the model is explaining more variance in the target variable, implying better performance Kolhatkar and Östblom (2023).

Reference

Kaufmann, Christoph C., Claudia Wegberger, Maximilian Tscharre, Paul M. Haller, Edita Piackova, Irena Vujasin, Mona Kassem, et al. 2020. “Effect of Marathon and Ultra-Marathon on Inflammation and Iron Homeostasis.” *Scandinavian Journal of Medicine & Science in Sports*. https://doi.org/10.1111/sms.13869.

Kolhatkar, V., and J. Östblom. 2023. “Regression Metrics [Lecture Notes].” UBC GitHub Pages. https://pages.github.ubc.ca/mds-2024-25/DSCI_573_feat-model-select_students/lectures/02_regression-metrics.html.

Taraji, Maryam, Paul R. Haddad, Ruth I. J. Amos, Mohammad Talebi, Roman Szucs, John W. Dolan, and Christopher A. Pohl. 2017. “Error Measures in Quantitative Structure-Retention Relationships Studies.” *Journal of Chromatography A* 1524: 298–302. https://doi.org/https://doi.org/10.1016/j.chroma.2017.09.050.

Zoladz, Jerzy A., and Zenon Nieckarz. 2021. “Marathon Race Performance Increases the Amount of Particulate Matter Deposited in the Respiratory System of Runners: An Incentive for ”Clean Air Marathon Runs”.” *PeerJ* 9: e11562. https://doi.org/10.7717/peerj.11562.