Project 3: Ironman

Justin Carter

4/6/2022

# Introduction

The purpose of this document is to utilize data from the 2020 and 2021 Ironman competitions to answer two research questions:

1. If and/or how performance metrics from the 2020 race subcomponents (swimming, biking, running, and transition time), along with demographic information, are useful for explaining overall performance in the 2021 race (in overall time)
2. If athletes that competed in both 2020 and 2021 were substantially different in terms of performance metrics (race subcomponents) than the athletes that only competed in 2021

The Data Exploration section and its subsections involve the reading-in of the datasets and combining them. As these datasets did not provide unique keys for each athlete between the competitions, the implications of joining the datasets were discussed. A scatter plot matrix was created in order to observe trends within and between variables in the dataset.

The Univariate Relationships with Overall 2021 Ironman Times section and its subsections encompass the creation of 4 simple linear regression models. These models utilized 2020 Ironman race subcomponent data in order to predict 2021 Ironman overall times. The results of the models were visualized and interpreted.

In the Multiple Linear Regression section (and its subsections), two similar multiple linear regression models were created in order to predict 2021 Ironman overall times utilizing multiple variables from the 2020 Ironman competition. Accuracy metrics between the models were compared, and a model was selected. The findings from the selected model were discussed.

Descriptive statistics between athletes that competed in both the 2020 and 2021 races and athletes that just competed in the 2020 race were displayed in the Descriptive Statistics section. Hypothesis tests were conducted in order to determine whether differences in performance between the two groups were statistically significant.

# Data Exploration

## Reading In Data and Transforming Times

Primarily, the data sets were read in, and the variables RunTime, SwimTime, BikeTime, and OverallTime were converted into numeric minutes from an hour-minute-second format. Since the format of SwimTime was variable based on whether or not an athlete completed the swimming portion in under or over an hour, it was assumed that values over 10 in the left-most time value were times that were under an hour. Times with values that were under 10 in the left-most time value for SwimTime were treated as times that were completed in the span of 1 or more hours. This was done in order to properly make computations with these variables.

## Creating the TransitionTime Variable

Transition time for each athlete was calculated as the difference of OverallTime and the sum of RunTime, SwimTime, and BikeTime.

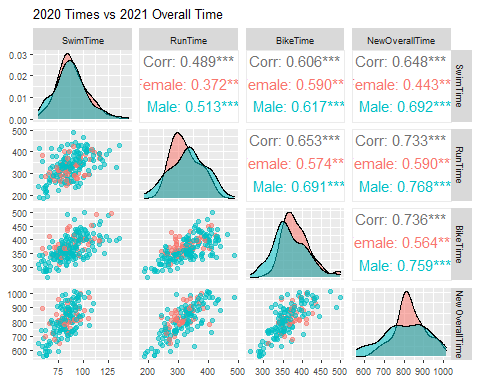
## Creating the AgeGroup Variable

A new variable, AgeGroup, was created for each athlete. This variable constituted a transformation of the Division variable. The gender attribute was removed, and AgeGroup was cleaned such that it only contained the age group of each athlete. If an athlete was labeled such that no age group was given (i.e., as MPRO, FPRO, MPC, or FPC), an NA value was assigned to their age group.

## Combining the Data Sets and Quality Check

An inner join was performed in order to combine the 2020 and 2021 data sets. The total number of rows between the two data sets before joining was 3687. 247 rows remained after performing this join. The vast majority of the remaining rows contained data for individuals that competed in both 2020 and 2021. A quality check was performed to better ensure the reliability of the data post-join. Rows where athletes had impossible differences in the AgeGroup variable between 2020 and 2021 were dropped. These rows likely contained information from different individuals that possessed the same name. Rows where the difference in age group could not be calculated were kept in the data set. After this cleaning was performed, 231 rows remained in the data set.

## Observe Relationships Between 2020 Times and 2021 Overall Time



To better understand the relationship between athletes’ performance in 2020 to their performance in 2021, their times in running, swimming and biking were plotted against each other, as well as against the overall time they took to complete the race in 2021. Correlations between these variables were calculated. As can be seen above, athletes’ performance in every venue were positively correlated with their overall time a year later, and moderate-to-high positive correlations were observed between the subcomponents of the race. SwimTime, RunTime, BikeTime, and NewOverallTime (overall time taken to complete the race in 2021) seem to approach normal distributions. The highest correlation with athletes’ overall time in 2021 (regardless of gender) was the time they took to complete the running portion of the race.

# Univariate Relationships with Overall 2021 Ironman Times

This section of the report covers univariate relationships between individual event times of swimming, biking, running, and transition time in 2020 to the variable of overall time in 2021.

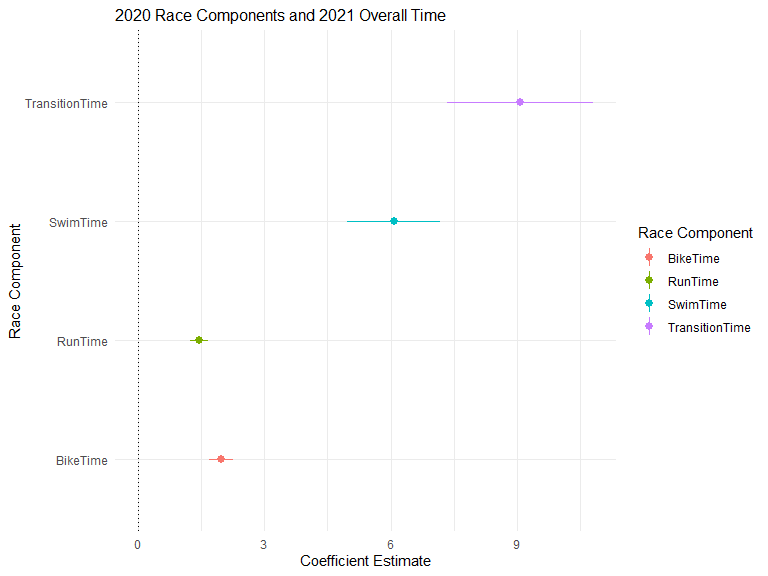
## Fitting the models

Four simple linear regression models were created to predict overall 2021 Ironman times. Each of the 2020 race components (swimming, biking, running, and transition times) were utilized to predict 2021 Ironman race times.

## Summary of Findings

The relationships that were measured by the simple regression models were summarized in a univariate regression table as well as a forest plot. Both can be found below.

| Univariate Regression | | | |
| --- | --- | --- | --- |
| Race Component | Estimate | 95% CI Lower Bound | 95% CI Upper Bound |
| SwimTime | 6.08 | 4.97 | 7.18 |
| BikeTime | 1.97 | 1.69 | 2.25 |
| RunTime | 1.45 | 1.24 | 1.66 |
| TransitionTime | 9.08 | 7.35 | 10.81 |



## Interpretation of Results

From the forest plot above, it appears that 2021 overall time changed the most per unit of change in 2020 transition time (9 minutes) compared to the other components of the 2020 Ironman. However, 2020 transition time also saw the most variability within its values compared to the other variables. All components of the 2020 Ironman were positively associated with overall time in 2021. In addition, all race components were found to have statistically-significant relationships with overall time at the p= .05 level.

# Multiple Linear Regression

Exploratory analysis was conducted in order to determine which variables in the data set were appropriate for fitting a multiple regression model to predict 2021 overall time in the Ironman.

The quantitative variables in the data set were correlated with each other in order to assess which variables were most associated with 2021 overall time, and which predictor variables were highly-associated with each other.

Predictor variables with correlations with other predictors at or above 0.7 were excluded from prospective models.

After removing said variables, two prospective multiple linear regression models were created.

Model 1 contained the variables:

* OverallTime
* SwimTime
* TransitionTime

Model 2 contained the variables:

* DivisionRank
* SwimTime
* BikeTime
* RunTime
* TransitionTime

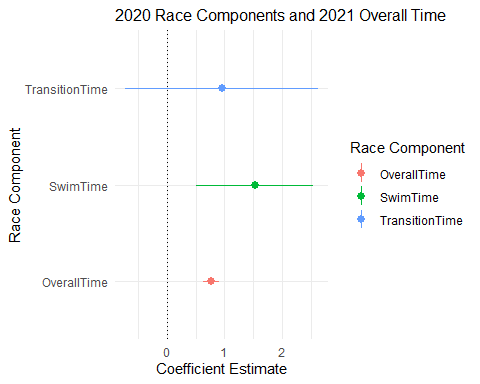
## Model Performance Assessment and Comparison

Models 1 and 2 were assessed via obtaining various error metrics. In order to do this, the data set was split into 2 groups: a test set (which encompassed 1/4 of the data) and a training set, which encompassed the remaining data. Both models were trained on the training set and then predicted values of 2021 overall time in the test set. Adjusted R Squared, AIC, MAPE, and RMSE were calculated for both models. The results can be seen below.

| Model Comparison | | | | |
| --- | --- | --- | --- | --- |
| Name | Adjusted R Squared | AIC | MAPE | RMSE |
| Model 1 | 0.72 | 1,751.47 | 4.77 | 44.59 |
| Model 2 | 0.73 | 1,749.59 | 4.76 | 44.48 |

As can be seen above, the error metrics between Model 1 and Model 2 are nearly identical. As Model 1 contained fewer variables and similar error metrics to the more complicated Model 2, Model 1 was selected as the final model to predict values of overall 2021 Ironman times. The results above suggest that Model 1’s predictions of 2021 overall time were off by 44.59 minutes on average. In addition, the model’s predictions were off by about 4.77% on average.

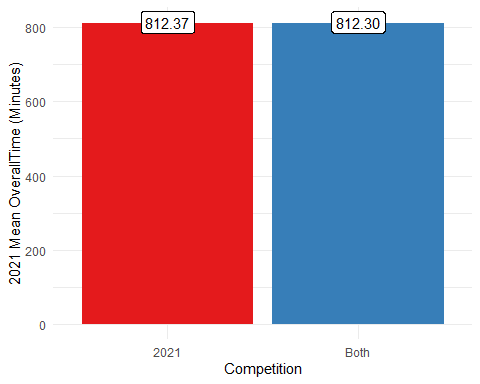
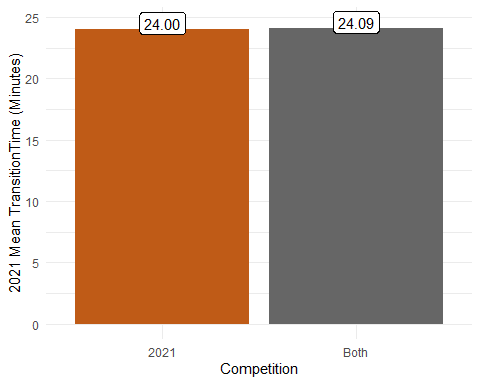
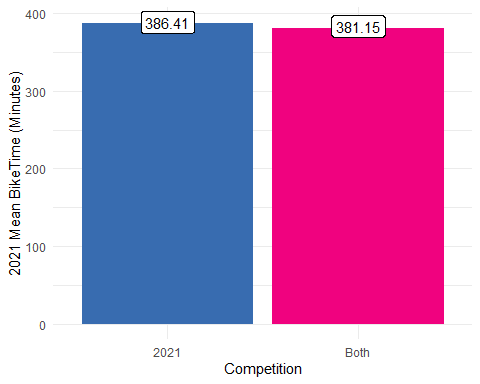
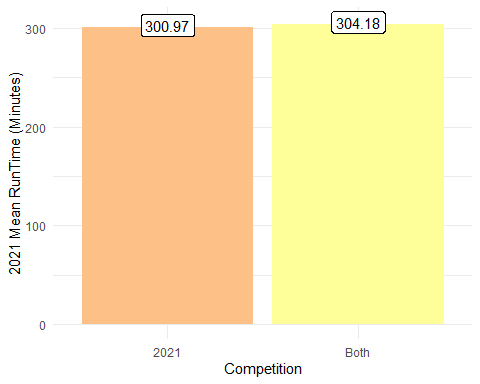
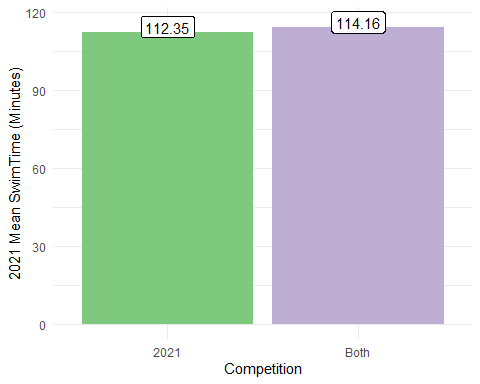
## Model Interpretation



From the forest plot above, it appears that 2021 overall time changed the most per unit of change in 2020 swim time (about 1.5 minutes) compared to the other components of the 2020 Ironman. 2020 transition time saw the most variability within its values compared to the other variables. All swim and overall times of the 2020 Ironman were positively associated with overall time in 2021. Transition time in 2020 had a coefficient estimate that was positive, however, its 95% confidence interval indicates that its relationship with 2021 overall time is questionable. Indeed, transition time in 2020 may have a negative relationship, or no relationship whatsoever to 2021 overall time. This finding, of course, starkly contrasts the results from the previous section. The distinction between these findings seems to be that when the other variables in the multiple regression model are held constant, transition time has little or no relationship with 2021 overall time. Race components (besides transition time) were found to have statistically-significant relationships with overall time at the p= .05 level.

# Descriptive Statistics

This section focuses on the comparison of athletes that competed in just the 2021 Ironman competition to athletes that competed in both the 2020 and 2021 Ironman.

Below, the mean times to complete each subcomponent of the 2021 Ironman were compared between athletes that competed in both the 2020 and 2021 Ironman race and athletes that just competed in the 2021 Ironman.

Welch Two Sample t-tests were conducted in order to determine whether or not the differences in mean times to complete each subcomponent of the 2021 Ironman competition were significant between the two groups of athletes. The results yielded no statistically significant differences in the means between the two groups of athletes in any subcomponent of the 2021 Ironman competition at the p = .05 level. That is to say, there is no evidence to suggest that there are significant differences in the performance of athletes that competed in both the 2020 and 2021 Ironman competitions compared to athletes that competed in just the 2021 Ironman.

## Conclusion

There are several key takeaways that can be drawn from the sections above. Primarily, 2020 transition time was best able to predict 2021 overall Ironman time (in terms of univariate models utilizing race subcomponents). It was found that the combination of 2020 swim time, 2020 overall time, and 2020 transition time was able to predict 2021 overall time within a 5% absolute margin of error. This error rate was similar to a more complicated model that utilized a larger number of variables. Finally, no statistical significance was observed between the differences of performances of athletes that competed in both Ironman competitions compared to athletes that competed in just the 2020 competition in any race subcomponent.