Analyzing Performance Metrics of the 2020 and 2021 Florida Ironman Race

Matt Nelson

2023-12-13

## I.Introduction

This project will consist of taking data from 2020 and 2021 Ironman race results to determine the following:

1. If and how performance metrics (run, bike, swim, and transition time) from the 2020 race are useful in explaining the overall performance in 2021.
2. Are athletes that competed in 2020 and 2021 substantially different in terms of performance metrics than the athletes that only competed in 2021.

We will begin the project by cleaning the 2020 and 2021 data sets, joining the data sets together, then running multiple quality checks to ensure there is a proper join. Taking the joined data set we will create a scatterplot matrix and begin to look for noteworthy trends between 2021 overall time and 2020 performance metrics along with gender.

Next we will investigate univariate relationships between individual times of the 2020 performance metrics and the 2021 overall time. This will be done by fitting each of the 4 2020 performance metrics into a simple linear regression models. We will take these 4 models and create a single univariate regression table and forest plot. This will help us get an understanding of how each performance metric individually impacts 2021 overall time.

As we begin to understand stand each variable’s influence, we can combine these into an explanantory multiple linear regression model to address the first research question. Our variables will be selected using the lasso regression function on our training data set to create a training model. The results will be measured by Adjusted , , and RMSE on the input of the test data into the training model.

We will address the 2nd research question by separating racers into 2 groups. One group will consist of racers who competed in both 2020 and 2021. The other being racers who only competed in 2021. We will compare these athletes and their performance metrics by creating overlapping density plots. This will help us identify time differences in any of the performance metrics based on these two groups.

## II. Relating 2020 and 2021 performance

In this portion of the project we will preprocess, join the 2020 and 2021 data sets, and filter out incorrectly recorded names from the joined data set. After this we will create a scatter plot matrix to analyze noteworthy trends between response and predictor variables.

**Preprocessing 2020 and 2021 Data sets** \* Convert swim, bike, run, and overall times to numeric minutes + Compute transition minutes + Remove Gender from the Division and create a new variable called AgeGroup - **Note:** PRO and PC Divisions will be removed from the data set, given the focus of Division is on age.

**Join 2020 and 2021 Data sets** We will join the 2020 and 2021 data sets by racer name. Being this is an improper join key, the joined data should be interpreted with caution due to potential for incomplete merging. Potential issues that could arise include things such misspelled name variations from 2020 to 2021, different racers having the same name, and names being unique to each data set not being included in the join.

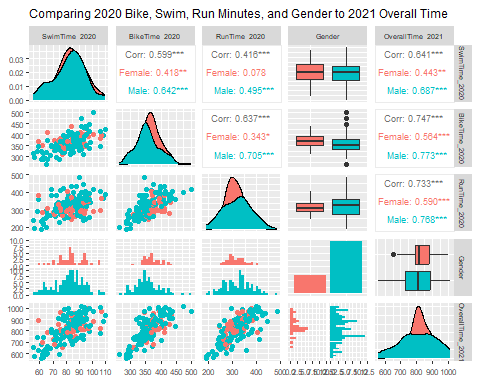
**Quality Check**

After searching for name matches that are likely not the same person, we see Kevin Young was in the 45-49 AgeGroup in 2020 and 30-34 AgeGroup in 2021. His name will be filtered out of the data set. In addition there are 2 John Briggs listed in the 2020 data set and only 1 in the 2021 data set. Since we are joining by Name, there is an incorrect join and his name will be filtered out as well.

## # A tibble: 1 × 3  
## Name AgeGroup\_2020 AgeGroup\_2021  
## <chr> <chr> <chr>   
## 1 Kevin Young 45-49 30-34

## # A tibble: 2 × 3  
## # Groups: Name [1]  
## Name AgeGroup\_2020 OverallTime\_2021  
## <chr> <chr> <dbl>  
## 1 John Briggs 50-54 800  
## 2 John Briggs 55-59 800

**Scatter plot matrix with Noteworthy trends** From looking at the scatter plot matrix below, all variables appear to be mostly normally distributed. There appears to be positive relationships between 2021 Overall minutes and 2020 Swim, Bike and Run minutes. These relationships all look to be linear. SwimTime has a larger variance in data points, whereas BikeTime and RunTime has a smaller variance, but this variance seems to slightly increase with time. There also appears to be a couple outliers as noted by the dots on the top end of the gender boxplots.

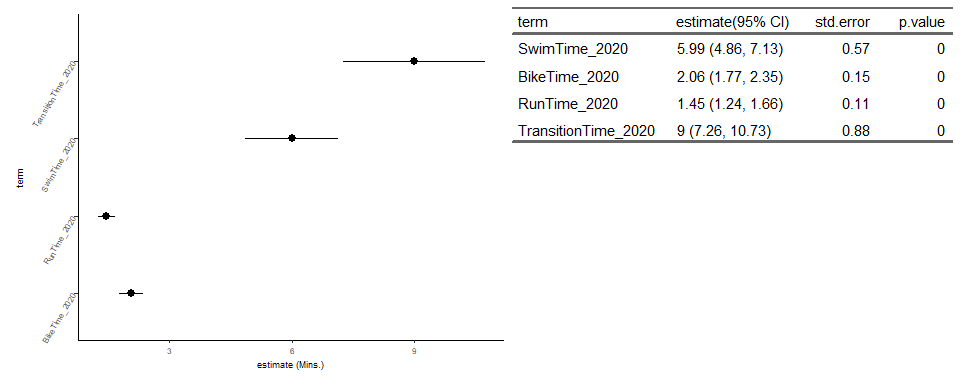


### A.Univariate Relationships

First we will fit four simple linear regression models predicting overall time in 2021; one model for each 2020 time metric (SwimTime, BikeTime, TransitionTime, and RunTime). This will allow us analyze how the effect of each of the 4 variables independently influence the OverallTime\_2021.

swim\_model <- lm(OverallTime\_2021 ~ SwimTime\_2020, data = check.filter)  
bike\_model <- lm(OverallTime\_2021 ~ BikeTime\_2020, data = check.filter)  
run\_model <- lm(OverallTime\_2021 ~ RunTime\_2020, data = check.filter)  
transition\_model <- lm(OverallTime\_2021 ~ TransitionTime\_2020,   
 data = check.filter)

**Univariate Analysis**



Combining these 4 models into a univariate regression table and forest plot above, we can see how each variable independently relates to the Overall 2021 time. All four variables are positively correlated with OverallTime\_2021 and are significant at a .05 confidence level. Even at the low end of the confidence interval a one minute increase in TransitionTime\_2020 has the largest predicted impact in OverallTime\_2021 increase. For every minute increase of 2020 TransitionTime, we can expect between a 7.26 and 10.73 minute increase in 2021 OverallTime.

### B.Multiple Regression Model

We want to determine if and how performance metrics from the 2020 race (swim time, bike time, run time, and transition time) along with demographic information (gender and age group) are useful for explaining overall performance in the 2021 race (overall time)

To do this we will first convert AgeGroup\_2020 to numeric, split our data into a training and test set, then input multiple combinations of predictor variables into a lasso regression model using the training data set. We will analyze the lasso selected variables to make sure they can be easily explained, especially focusing on any interaction terms. Then we will input our test data into the trained lasso model and look for maximized Adj. and values, along with minimized RMSE values of the test data. We will repeat the process until there is a satisfactory model that is easily interpreted and still fits the data well.

After several iterations we will be using the following linear model:

Y = 111.78 + 1.81 + .71 + .78 + 2.08 - 24.87

**Analysis**

Our multiple linear regression model shows all 4 2020 performance components having an impact on the 2021 Overall time. This can be seen on the Variable Coefficients table below. 2020 SwimTime is the most impactful on the 2021 Overall Time of these 4. On average, for every minute increase in SwimTime, we can expect a 2.04 minute increase in OverallTime. For every minute increase in TransitionTime we can expect 1.81 minute increase, BikeTime a .71 minute increase, and RunTime as .78 minute increase. In addition, we can expect a competing male to perform approximately 25 minutes better than a female, when all other variables are fixed.

Our value of .713 on the test data set tells us we can expect our regression line to account for 71% of the variance. In addition we see an RMSE value of 60.95.

Variable Coefficients from Lasso Regression Using Optimal Lambda

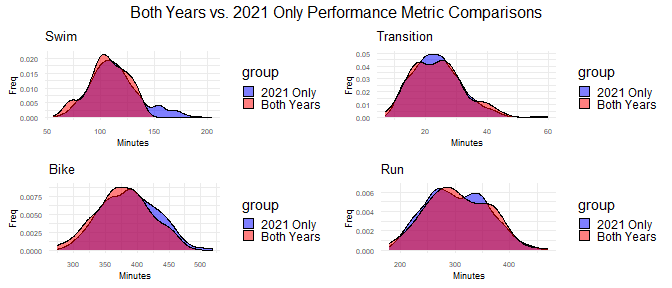
| Variable | Coefficient |
| --- | --- |
| TransitionTime\_2020 | 1.81 |
| BikeTime\_2020 | 0.71 |
| RunTime\_2020 | 0.78 |
| SwimTime\_2020 | 2.09 |
| GenderMale | -24.87 |

Statisical Metrics

| RMSE | Adj R^2 | R^2 |
| --- | --- | --- |
| 60.947 | 0.678 | 0.713 |

## III.How Repeat Athletes Compare To Others

Here we will compare performance metrics of athletes that competed in both 2020 and 2021 to those that competed in 2021 only. We will do this by creating a data set with the 2 groups. One being athletes that competed in both years. The other being athletes that competed in only 2021. We will create overlapping density plots of each performance metric. This should help highlight differences in performance metrics between the two groups.



| Both Years | | | Only 2021 | |
| --- | --- | --- | --- | --- |
| Performance.Metric | Mean | StDev | Mean. | StDev. |
| 2021 SwimTime | 106.98 | 18.17 | 113.76 | 23.62 |
| 2021 RunTime | 304.67 | 56.05 | 303.38 | 55.59 |
| 2021 BikeTime | 379.37 | 42.23 | 388.69 | 43.65 |
| 2021 TransitionTime | 23.79 | 8.18 | 23.81 | 8.11 |

**Analysis** From looking at the overlapping density plots, swim times appear to have the largest difference in distribution between ‘2021 only’ racers and ‘both years’ racers. This difference is most notable at the tail end. There appears to be a larger proportion of racers’ swim times between 150-200 minutes in the 2021 only group. Bike time look to have a similar distribution between the 2 groups, however, the average time looks to be lower for the the racers who competed in both years. Run time and transition time each seem to have similar average times along with a similar distribution pattern between the 2 racer groups.

Note: This analysis is strictly a comparison of performance metrics and how they compare between the 2 groups. Factors such as gender, and age group are not included in this analysis.