Project 3

Jeffrey Orcutt

2023-11-15

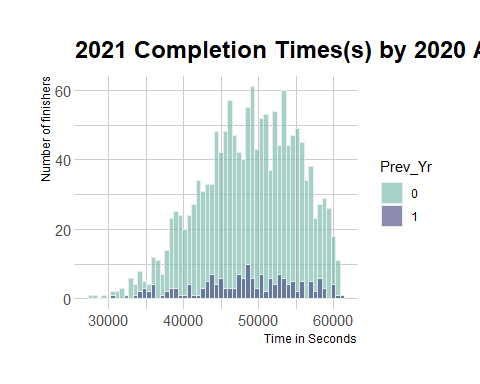
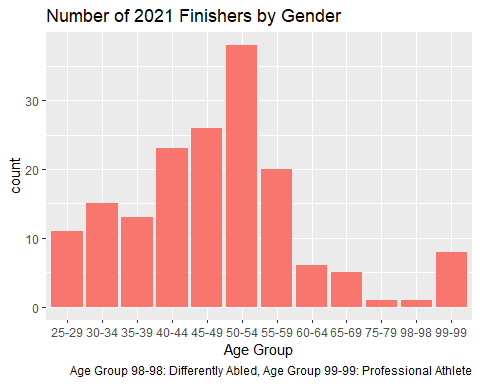
#Introduction Problem Definition: Given two datasets from Florida Ironman events in 2020 and 2021, we are challenged to answer two research questions: First, determine what and to what extent, if any, participants’ 2020 performance data help explain their 2021 overall performance. Secondarily, establish whether those participants who participated in both races performed better than those who only participated in 2021.

##Data introduction: The datasets include observations of participants’ names, demographics, event performance, and completion status. In 2020, 1354 participants were registered: 113 did not start, 99 did not finish, 2 were disqualified and 1140 finished the race. In 2021, 2333 participants were registered: 71 did not start, 563 did not finish, 2 were disqualified and 1697 finished the race. In 2021, there is a broad range of overall finish times ranging from 7:42:57 to 16:57:23.

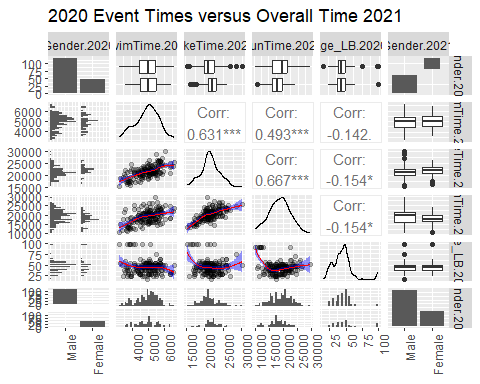
##Data Preparation: Both problems ask directly about people who finished the race, I’ll begin by removing those that did not start, finish, or were disqualified from either race. For simplicity’s sake, I standardized all times into HH:MM:SS format. Divisional age groups were transformed into factors of their age group with two special groups being added: 98-98 encoding is the differently able-bodied competitors, and professionals were encoded as 99-99 age group. For further exploration, I created some time difference values from each of the events, overall, and calculated the difference in transition times - the time spent transitioning from one event to the next.

##Matching Data to Problems:  
Because the datasets are being joined with the Name value as their primary keys, as such, some names may match but not be the same person in both years. Some validation was done to check the matched participant’s gender and age but it is possible that error was introduced into this analysis at this point. Of the 170 participants who matched by names only, only 1 participant was removed because of age group mismatches. Additionally, I performed checks for an extreme increase in overall race time of 3 hours or more, and a decrease of 2 hours or more. While some records were identified as fulfilling these criteria, none were removed from the final dataset. Lastly, changes in gender from year to year with none identified.

I ended the preparation steps with two sets of data: one that will be used to investigate the previous year’s event times on the following year’s overall time and another dataset that has all the finishers from 2021 data with one additional column denoting whether that person also finished the previous year’s race or not to investigate if there is a statistically significant difference in the performances of finishers who competed in both years versus those that only competed in 2021.

I created two exploratory visualizations for this portion of the project to become more familiar with the data. The visualization on the left shows the age distribution of all finishers in 2021. The visualization on the right describes the distribution of overall finish time by previous participation, where green is the people who have participated in both 2020 and 2021.  


Lastly, I created a plot to pairwise look at the individual event performances in 2020 of those who completed both races against their performance in 2021 to see if there were correlations. It would appear that there is a statistically significant correlation for all three events against 2021 performance and that relationship is likely positively correlated. All four distributions of time have a roughly normal distribution and the three event times have a positive, roughly linear trend line when graphed against the overall finish time.



Outliers exist in the Bike and Run Times when viewed by gender. Several males took more than 7:30:00 in the bicycling event while a few completed the event in 5:00:00. On the female side of the bicycling event, several finished in less than 6:15:00. In the running event, one male finished in 8:07:10 to complete the event and six females finished in less than 3:40:00. Four females finished the entire race in under 10:00:00. Overall, I would expect that the longer it took participants to complete each of the three events in 2020 that their 2021 overall time would be longer as well.

#Univariate Analysis  
When I performed univariate analysis on assessing the individual effects of individual event times from 20201 on overall performance in 2021, I found a couple of effects that are worth discussing. To begin this process, I fitted four models one for each of the three 2021 event times and the time spent transitioning between events in 2021 against the response variable of the 2022 overall time. All four models consisting of individual times produced models that were statistically significant and all were positive relationships. As a validity check, I took the step of checking for multi-colinearity and found that while there are correlations, none of the individual features would be considered highly correlated with another feature having a correlation value of .9 or more.

Firstly, the largest individual effect wasn’t even an event, but rather the amount of time it takes to transition from event to event. For every second spent transitioning, one could, with 95% confidence, estimate that they are adding between 8.3 and 12 seconds to their overall time. This makes sense because the time between events is essentially non-productive time and participants have a wide variety of experiences and different goals when choosing to participate: some are there to win, some are there to simply finish. Additionally, transition time may be used to take care of pressing needs such as eating, injuries, and other self-care needs.  
The second largest individual effect was the swim time. When viewed by itself, one would expect every second spent swimming to have an effect on overall time in 2021 of between 5.6 to 7.8 seconds. A participant’s swimming time appears to explain a significant portion of their overall performance, as it is the first event participants encounter and generally regarded as the most difficult so a longer performance in this event indicates that a participant is going to have a disproportionately longer overall time.

Lastly, all four features were individually statistically significant in this univariate analysis. All four features indicated a larger than one-to-one ratio of seconds per event to overall performance as individual features. I anticipate that a participant’s overall time will be more affected by their swim and transition times than their run and bike times.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Characteristic** | **Beta** | **95% CI**1 | **p-value** | | --- | --- | --- | --- | | SwimTime.2020 | 6.7 | 5.6, 7.8 | <0.001 | | BikeTime.2020 | 2.2 | 1.9, 2.5 | <0.001 | | RunTime.2020 | 1.5 | 1.4, 1.7 | <0.001 | | TransitionTime.2020 | 10 | 8.3, 12 | <0.001 | | 1CI = Confidence Interval | | | | |  |

#Multiple Regression Models I created six multiple regression models to investigate the ability of 2020 event times and demographic combinations to assess how well they compare with each other in predicting 2021 times. These models all had different combinations of the features to try to find the model that best explains the relationships between 2020 events and 2021 performance.

The end goal of this process of trial and error with modeling is to create an interpretable model that is explanatory in the relationship between the features and the response variable of interest. This requires two things in my opinion, a model that can be explained in detail to those with less knowledge and that the model can accurately predict the variable of interest. The features contained in the model generally need to be statistically significant.

In comparing the models to one another, two standard methods include comparing values known as the adjusted R squared and the Akaike Information Criterion (AIC). For our models, the Adjusted R Squared value explains how much of the variability in their 2021 Overall Time is explained by the features of the model. In adjusted r squared, the values range between zero and one, where one is a perfect fit - therefore higher is better. AIC is simply a metric that compares the goodness-of-fit of the model against the data while penalizing the model for the number of model coefficients and helps prevent overfitting.

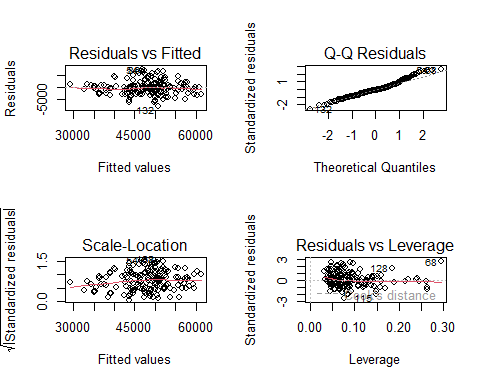
| row\_names | All.Predictors | Performance.Events.and.Age.Division | All.Times.and.Age.Division | Age.Bands.and.All.Predictors | All.Times | Event.Times.with.Gender.Interaction |
| --- | --- | --- | --- | --- | --- | --- |
| Attributes | Age\_Group, Gender, SwimTime, BikeTime, RunTime, and TransitionTime | SwimTime, BikeTime, RunTime, and Age\_Group | SwimTime, BikeTime, RunTime, TransitionTime, and Age\_Group | Gender, age\_band, SwimTime, BikeTime, RunTime, and TransitionTime | SwimTime, BikeTime, RunTime, and TransitionTime | Gender, SwimTime, BikeTime, RunTime, and TransitionTime |
| AIC | 2700.6829 | 2705.4501 | 2703.929 | 2712.3239 | 2721.0328 | 2726.816 |
| Adjusted R^2 | 0.8044 | 0.7966 | 0.7995 | 0.7869 | 0.7625 | 0.7611 |
| Subjective Rank | 1 | 2 | 3 | 4 | 5 | 6 |
| Rejection Reason | None | None | Transition Time not Stat Sig. | Age Band not Stat Sig. | Transition Time not Stat Sig. | No interaction Stat Sig. |

In the table above, I’ve subjectively rank-ordered what I feel are the best models for the dataset. I’ve also included in the table the rejection reason for each model not discussed below.  
In choosing between the top two models, this is where the AIC and Adjusted R^2 values came into play. Based solely on simplicity, the model ranked number 2 should win, however, based on Burnham & Anderson, 2004 – the difference in the two models’ AIC values indicates that the model containing just performance events and age division is significantly less likely to be a better model than the one containing all predictors. Perhaps, with a larger dataset, we would find that overfitting has become a problem and the second-ranked model, or even potentially another model, would be a better description of the relationship.

The best model’s features are shown below:

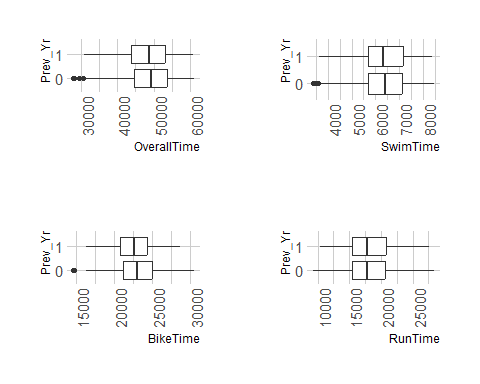
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 17697.7782958 4096.4955629 4.320224 2.833603e-05  
## Age\_Group.202025-29 -13429.3475175 3234.9428276 -4.151340 5.540425e-05  
## Age\_Group.202030-34 -13182.1163179 3180.4044161 -4.144792 5.684198e-05  
## Age\_Group.202035-39 -13007.2029171 3237.9314557 -4.017133 9.313658e-05  
## Age\_Group.202040-44 -14149.3047898 3162.9339692 -4.473475 1.517661e-05  
## Age\_Group.202045-49 -11328.1952018 3158.3756867 -3.586716 4.534073e-04  
## Age\_Group.202050-54 -12403.7653678 3142.6964977 -3.946854 1.216601e-04  
## Age\_Group.202055-59 -11216.8175448 3177.4396287 -3.530143 5.528767e-04  
## Age\_Group.202060-64 -11032.9551064 3403.5600290 -3.241593 1.466592e-03  
## Age\_Group.202065-69 -9051.4580528 3479.0298597 -2.601719 1.021165e-02  
## Age\_Group.202075-79 -10161.3683603 4450.5266114 -2.283183 2.383370e-02  
## Age\_Group.202098-98 -5520.7019360 4421.8271813 -1.248511 2.138020e-01  
## Age\_Group.202099-99 -15115.9310566 3349.1269365 -4.513394 1.286708e-05  
## Gender.2020Female 1268.2749055 581.6241046 2.180575 3.078388e-02  
## SwimTime.2020 2.4408868 0.4707702 5.184880 6.934068e-07  
## BikeTime.2020 0.6340493 0.1731499 3.661852 3.471758e-04  
## RunTime.2020 0.7896824 0.1057749 7.465688 6.428315e-12  
## TransitionTime.2020 1.5507793 0.7559960 2.051306 4.198851e-02

Starting with the intercept, you can see that the model estimate starts at 17,697.78 seconds. This means that since there are two factors, the base case for the intercept is a male between 18-24 years old. Being female is expected to add 1268.27 seconds to a participant’s overall score. The age group factor serves as an offset to the intercept base group of 18-24, based on a participant’s age to their overall time. The offset generally decreases as age increases, with professionals getting the largest offset of -15,115.93 seconds and 65-69-year-olds getting the smallest significant offset of -9051.46 seconds. The four predicting times are located towards the bottom of the list, notice that Swim Time adds 2.4409 seconds per second in the event and every second spent transitioning between events adds 1.5508 seconds to the expected overall time of a given participant. The contribution to overall time between the bike event time is 0.634 seconds per second in the event and 0.7897 in the run event.



When applied to the whole dataset, the model has a mean absolute percentage error (MAPE) of 4.61% and a Root Mean Squared Error of 2916.36. A MAPE of less than 5% is considered to be a model that can sufficiently explain the relationship between the value of interest and the features of the model. Lastly, as a validation, I’ve displayed below the diagnostic four-pack of plots which do not show any cause for concern of the model’s validity. Overall, this model fulfills our requirement to have a reasonably easily explained model that explains how 2020 times relate to 2021 performance.

#Secondary Research Question  
In our second research question, we were asked to investigate whether there is a statistical difference in the performance of athletes during each event who participated in both years. Earlier in this project, I showed the distribution of 2021 overall times based on 2020 participation or not. Below you’ll find a four pack of box charts showing the distribution of times by events in 2021 based on the participants previous year participation. These distributions show that there is an appearance of lower times in most of the times. I did not do any significance testing for the populations, so it is unknown whether the apparent relationships between the populations would be statistically significant.



#Summary  
For the primary research question, we were able to create a model that explains approximately 82.4% of the variance in Overall time in 2021 with 2020 event performance times with a model that can be explained in its entirety in a single paragraph. Furthermore, this model is able to predict an outcome in 2021 given 2020 event times with a MAPE of 4.61%. The secondary research question appears to have some relationships that indicate that participants who participated in both years performed, on average, better than those who did not.