Final Project 3

Adam Gruber

2022-12-07

Intro

Our Goal is to predict the second year performance of Triathlon athletes based on their times in 2020. This required several transformations for the data to find all the people that participated in both years as well as taking the data from being granular to ready to use. There are considerable more partcipants in the year 2021 compared to 2020. 1680 vs 1133. That is a 48% increase. This is likely due to poor turn out from Covid 19. Considering this study needs people to repeat the Triathlon in order to make test and make accurate predictions, we will likely need to repeat this study using 2022 and 2023 in order to build a proper model. Our best refined model shows adjusted r squared of .766 for predicting the overall times. Our secondary goal of determining if there is an improvement in overall times when comparing first time and repeat participants is inconclusive, showing little time differences between the groups.

I created a separate age group variable. This variable was created from the Division Variable. This Variable lacks the age for people who were Professional competitors. This means we are lacking age for professionals, but these could be replaced with the median, or mode. A table of our participants shows the age skews to the higher end. If we used mode, the age of professionals would be replaced with 50-54 category. This seems a bit unrealistic to expect most of the professionals to be a bit older. Natural inclination is professionals would be young and at peak fitness, but without further proof of age I choose to remove the 33 professionals from the 2020 data set. PC means physically challenged and those 4 were removed as well for similar reasons from 2020 Raw Data Set. The 2021 Raw data set had 5 unknown. Those removed as well. This was the only way to ensure our data was as accurate as possible.

Joining the Data Frames

The data from 2020 triathlon and 2021 were combined together in one data frame. The data was joined by using the name of the participants, assuming name did not change. We were unable to use other variables to join the data set because they also change. For example, bib number, country, or rank. For example there are 6 participants in our refined data set that changed country. All of them were recorded as having moved to the USA. This seems it could have possibly been poor data collection or all of them did come from outside the USA and moved in the past year. Using the Name could potentially lead to issues since people might legally change their name for a variety of reasons such as : marriage, divorce, changing self identification.

Another example is changing ages. Inaccurate measurements or natural aging process can cause a change in the Age category. I checked the ages for our participants. I was looking for any people that aged backwards since this likely represented clerical errors that could lead to inaccurate regression. I found “Kevin Young”, and he was removed.

Our goal is to predict the 2021 performance of participants based on their times in 2020. I graphed this using ggpairs to show what linear correlation there might be in various scatter plots.

Looking at the overall differences in general, the Overall time for most athletes that completed both triathlons were similar. This may show a slight improvement over time but the mean time increased from 13 Hours 7 Minutes to 13 Hours 35 Minutes. The graphs Bike time and run time have a positive slope when graphed against Overall Time from 2021, showing at least a positive relationship. There appears to be a strong correlation between Bike Time in 2020,Run Time in 2020 and 2021 over all times. A correlation means at .743 of the Overall time can be explained by the Bike Time in 2020.

When comparing the genders the female graph is more normal compared to the male graph. The male chart has more outliers on both ends.

Overall Time

| Year | Min | Median | Mean | Max |
| --- | --- | --- | --- | --- |
| 2020 | 9 H 1 M | 13 H 10 M | 13 H 7 M | 17 H 22 M |
| 2021 | 9 H 18 M | 13 H 34 M | 13 H 35 M | 16 H 55 M |

Univariate Models

To predict the overall time for 2021 we created 4 prediction models based on the time for 2020. One based on each of the following : Swim Time, Run Time, Bike Time, Transition Time.

Each model had the P value listed as a significant term, which there was only one deciding variable for the model. Each Model varied in its R squared value. R squared shows the the proportion of Overall time that can be explained by the model. The bike model shows the highest r squared value and that it best explains the 2021 Overall Times.

When building a model that uses all 4 variables(All\_Variables Model), it has the highest R squared value, of 0.7168. This indicates that any one variable is insufficient to properly explain the 2021 over all times. This likely comes from participants having strengths and weaknesses in each event. Each event requires different skill sets. SO taking the performance of any one person in one event can can be unreliable when trying to predict their total time for 3 events. Predicting their overall time makes more sense when factoring in all events.

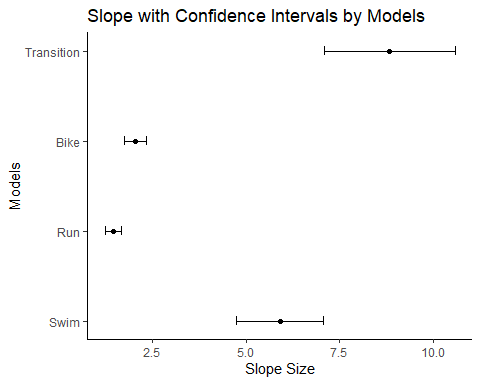
The All variables model was also tested to see if Gender was a significant factor for a linear model. It was not. The associated P value for gender was 0.119442 which means it is not significant for predicting/explaining the 2021 Overall Times.

R Squared Comparision

| Swim\_model | Run\_model | Bike\_model | Transition\_model | All\_Variables |
| --- | --- | --- | --- | --- |
| 0.3948779 | 0.5275633 | 0.5519581 | 0.3887383 | 0.71681 |

Confidence Intervals by Model

| Models | Index | Slope | Lower | Upper |
| --- | --- | --- | --- | --- |
| Swim | 1 | 5.897700 | 4.743082 | 7.052335 |
| Run | 2 | 1.434142 | 1.219510 | 1.648774 |
| Bike | 3 | 2.037950 | 1.747571 | 2.328334 |
| Transition | 4 | 8.829000 | 7.078081 | 10.579900 |

 The Confidence intervals for slope of Swim Time, and Transition time are much wider than the Confidence Intervals for Bike and Run Time. This is reflected in r squared values as well. All the r squared values are positive with bike and run showing the strongest correlation. The Transition Model and the Swim Model have very steep slopes, meaning as for every one minute increase in transition time, there is a 8.829 increase in overall time. This can be reflective of the overall size of transition time being smaller. The difference between max and min is 35.1 minutes. Difference between max and min run time is 299 minutes. Difference between max and min overall time is 457 minutes, and this means the difference 35 minute difference in transition time is trying to explain a 457 minute difference in over all time. Compare that to Run time that has a time difference of 299 minutes trying to explain the 457 minute difference in overall time. The slope will be more gradual for each unit increase in run time.

Using Single Variable models, our current best model is the Bike model.

Multivariate models

When adding Age group as a factor it can vary wildly for how it will interact with prediction model. Age is a factorial variable with 11 potential levels. Since we are looking at only those that had an overall time for both 2021 and 2020, there are not participants in all the age groups. Age as a variable has a negative coefficient as the age continues to increase.

To find the best possible model, created several different models and then compared them.

The simple model predicts Overall\_time\_21 based on Transition\_time\_20 + swim\_min\_20+ BikeTime\_20+ RunTime\_20 + AgeGroup.

The summary showed all variables were significant based on their P-value.

The interaction model used the same variables but looked at each variable and compared it against each age group.The summary of this model checks each age group if they are significant based on their P-value. After checking none of the variables were considered significant by themselves since their were 60 + variables.

The refined model was created using the forward step wise function . Step wise created the best model based on AIC. The stepwise function started at a null model that predicted Overall Time 21 based on the increase of 1. The Stepwise then added variables until it arrived at

Overall\_time\_21 predicted by BikeTime\_20 + RunTime\_20 + swim\_min\_20 + AgeGroup + Transition\_time\_20 + BikeTime\_20:swim\_min\_20 + RunTime\_20:AgeGroup

The 3 models were then compared based on their Adjusted R squared, AIC, and BIC.

Models for Overall time using Age Group

| Models | adj.r.sq | AIC | BIC |
| --- | --- | --- | --- |
| Simple | 0.7496466 | 1719.492 | 1768.493 |
| Interactive | 0.7505418 | 1745.620 | 1907.937 |
| Refined | 0.7660510 | 1716.513 | 1793.078 |

The simple model has the best BIC since it has the lowest. The refined model has the best AIC, of course since its the lowest. It also has the highest adjusted r squared value of .766051 meaning it has the best correlation to explain the the overall time for 2021. Many of the variables are considered significant. Runtime by itself is not considered significant, but when runtime interacts with Age group, it produces all significant terms in this model except for the 25-29 bracket. Gender was not chosen for the refined model as it increased the AIC score by adding additional noise to the model.

Run time when it interacts with Age group has a very small confidence interval for each term. Each of those terms was significant in the model. Age group by itself can have wide confidence interval for the slope. This could be explained by a variety of factors such as the changing health, lifestyle, and income of the participants. The amount of time and experience they have dedicated to practicing for the Triathlon can play large factors. Income generally increases as people age. Income was not measured but would be an interesting predictor variable of performance.

The best model to predict is:

Overall\_time\_21 ~ Bike Time\_20 + Run Time\_20 + swim\_min\_20 + Age Group + Transition\_time\_20 + Bike Time\_20:swim\_min\_20 + Run Time\_20:Age Group

The best model to simply explain the relationship without much loss in confidence was the Simple model

Overall\_time\_21 ~ Transition\_time\_20 + swim\_min\_20+ Bike Time\_20+ Run Time\_20 + Age Group

The difference in AIC is 3. A minimal loss in prediction for increased explain ability.

The best way to reduce times compared to the reference group of age (18 - 24) is to increase in age groups. The other major variables had positive slope. Since transition time and swim times in 2020 had the largest coefficient, and a poor time in 2020 would indicate a greater overall time in 2021. They would be categories a participant could focus on reducing in order to improve in time and potentially rank. Coefficient for Transition time and swim time are still both very small in the simple model: 1.3808 and 2.3181 respectively. They are small compared to coefficients like age group for example (25 - 29) has -218.7131 coefficient compared to (18 -24) reference group. Overall all the categories from 2020 Triathlon have a positive relationship with over all time from 2021. As times increased in 2020, Overall 2021 time increased.

**Does repeat experience help predict the future times?**

When comparing the Overall times for those that competed both years and those that only competed in 2021, they are very similar. The mean overall time for repeats was 815 minutes. The mean overall for non repeats was 818 minutes. The max, min, and median were both very similar on every level. This includes Run time, Swim Time, and Bike Time. The biggest difference appears to in max times.They are slightly smaller in many of the categories like Transition Time, or bike time. Max for non repeats was 498 minutes. Max for repeats was 477 minutes. The repeating the triathlon does not appear to improve the overall time significantly. This does not take into a account, the overall number of participants for 2021 is much larger than 2020. 1680 people in 2021 vs 1133 people in 2020. That is a 48 % increase. There could be many people that took off the 2020 triathlon due to events like Covid. This could lead to many former participants joining back in for 2021. This would lead to it appearing that many first time participants are quite skilled and have similar times to those repeat participants. The real solution to this would be looking back at older data from years past where there were similar number of participants or looking at future data.

## [1] "Multi Year Participants: Transition Time"

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 7.217 16.938 22.867 23.189 28.342 44.950

## [1] "2021 only Partcipants : Transition Time"

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 7.817 17.500 22.767 23.321 28.150 58.767

## [1] "Multi Year Participants: Swim Time"

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 60.68 96.37 105.84 106.94 120.10 139.47

## [1] "2021 only Partcipants : Swim Time"

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 56.28 95.73 107.22 106.59 118.70 140.82

Summary

When trying to determine explain 2021 times it was determined the uni-variate models had very little correlation individually and did not do an adequate job of explaining the overall times. The The multi-variate model was better able to explain the overall time for 2021 based on the 2020 times. The simple model had an adjusted r squared of .74 which was higher than any of the single variate models. This makes sense since no one activity ultimately determines the overall success. Each person may have strengths and weaknesses at each sport and this is reflected when the multi-variate model is more accurate at predicting.

There were two multivariate models that were determined to be useful. The simple model was more useful to explain the relationship between each variable. The refined model was better for actual predictions. There was only a minimal loss in correlation in r squared, Bic, or AIC. The refined model could be used for more accurate calculations when making predictions on individual competitors.

The results show the best way to reduce Overall time for participants in the 18 -24 age group is to increase in age groups until reaching the 40 -44 age group when model shows the age group coefficient becomes less impactful.

Using our limited data from only two years, it appears we are able to predict the Overall time for participants when they have competed previously. When comparing our overall times, and individual event times for multi year participants and single year participants, there appears to be little variation in times when comparing max, min,mean, and median. Our current model works to predict the overall times for multi year participants but it can be improved with more data and a larger population size. Covid potentially skewed the numbers in unforseen ways and looking at a longer history of races may show more varation in Overall times and the models.