Brett Kelly

DAT 490

Table of Contents

Abstract/Executive Summary

Long distance triathlons combine 3 sports into one continuous race. Running a good race requires a lot of planning and preparation. By using sufficient data and effective filtering measures one can make accurate predictions about the race depending on ones' goals. The specific race distance is the 70.3-mile distance. Results were gathered from Ironman's website. Various statistical analyses were conducted on the data. The results of the data showed that important variables such as age, weight, gender and being a good cyclist were important. In conclusion, athletes considering reaching their peak race potential should optimize for these variables to become truly legendary.

Project Plan

Project Plan - Overview of your Capstone project which should include a description of the data, proposed research question(s) and proposed methodology.

Data - Include the website and a brief description (one paragraph) of the data and potential variables.

Research Questions - Include 3 possible research questions with a brief description (one paragraph) regarding the objective for answering each question.

Methodology - What statistical procedures did you use to analyze the data?

The data I will be using is located here. https://www.kaggle.com/datasets/aiaiaidavid/ironman-703-race-data-between-2004-and-2020

The data is from the Ironman official website. It includes results from 2004 to 2020 from 70.3 length races. This includes a breakdown of each swim, bike and run split, as well as time taken in transition.

3 possible questions

1. Which of the 3 disciplines has a larger impact on overall

placement?

A tougher question than you would think. Each segment is of a different length and each segment comes after the other one. Fatigue in one segment carries over to the next one, and so on. I anticipate running, which is the last event, will matter greatly when it is hot, as the accumulation of fatigue will be a large factor at that time.

2. Which gender is faster in transition?

With a large amount of data, I should be able to see if gender plays a role in transition and if there is, try to find data that could support a reason for the difference.

3. What would it take to qualify for Ironman 70.3 World Championship?

I should simplify this question because qualifying is surprisingly complicated. For example, if finishers refuse slots, then a rolldown occurs and things can get strange in a hurry. For this question I will assume that every slot will get taken. I will assume the amount of qualifying spots is consistent. Also, I will exclude races within some threshold that are deemed to be too fast or too slow a course. Course variables such as altitude, hills, turns, etc. all make courses very different from one another. By looking at the final qualifying spots I should be able to find a range of splits and transition times that give higher and higher probabilities of qualifying.

Literature Review

Ironman triathlon is a sport consisting of a swim, bike and run components. I will be focusing on the 70.3-mile race distance but there are other versions that are either shorter or longer than this. This is a test of endurance that goes far beyond most types of racing. Cardiovascular, muscular, digestive systems, to name a few, are all pushed very hard over the course of 4-8 hours in this type of endurance sport. Other than the race considerations, one must consider training methodologies, nutrition and equipment. For the purpose of this review, I will identify optimizations that would help in qualifying for the Ironman World Championship while giving a general overview of the topic.

An article by the Institute of Primary Care, University of Zurich, states that, “Elite triathletes are generally tall, of average to light weight, and have low levels of body fat. Several studies investigated the changes in body composition during an ultra-triathlon and potential associations between anthropometric characteristics with split and overall race times.” (Knechtle, 153). Body mass index is a decent way of calculating an ideal body composition. A BMI of 21 is a good target for a triathlete. Any lower, would risk not having enough additional energy stores, such as fat, for the end of the race. Any higher, and the impact forces during the run would be too high. A well-balanced diet of carbohydrates and protein is necessary to maintain a healthy weight to height ratio.

The training schedule should be periodized. Meaning that some days should be hard efforts, followed by easy days. There should be rest days and rest weeks to allow training adaptation to happen. Too much intensity for too long can lead to chronic fatigue and poor performance. Time spent on each

sport should be balanced in a way to imitate the race itself. The swim takes approximately 30-45 minutes, the bike 2-3 hours and the run 1-2 hours depending on many factors.

Swimming requires swimming trunks, goggles, and possibly a wetsuit to deal with cold water. The bike portion requires a bike and helmet. Aerodynamics play a large part on the bike. Everything on the bike can be lighter and more aero to gain speed advantages. Even the clothes you wear make you faster. Running requires running shoes. There are a lot of non-essential items beyond this that can make the race more comfortable and make you faster.

A race should be raced at even pace to produce a good result. Many factors influence pacing, such as temperature, terrain, humidity and wind. This is even more challenging considering that there are 3 events to do, one right after the other. This ties directly into the first research question, which of the 3 disciplines has a larger impact on overall placement? The swim is first, and over the quickest. The bike is longest portion of the race, in both length and duration. The run is at the end of the race, when the body has expended most of its' energy.

"For Ironman triathletes, dehydration has been reported as a cause of fatigue, and exercise-associated hyponatremia has been highlighted as a major concern during such races." (Knechtle, 286). Most of our bodies are made up of water. So naturally, after spending many hours sweating, your body needs to top off on its water supply. This is difficult though, because your body can only process about 8 fluid ounces every 15 minutes. This means that to avoid dehydration, one must constantly be intaking small amounts of water. Too much water will slosh around and make you sick. Energy drinks are helpful in this process by keeping your sodium and electrolyte levels balanced so your body can properly process the water intake.

An article by the Institute of Primary Care, University of Zurich, states that, "Athletes competing in an Ironman ingest 3,643 kcal and expend 11,009±664 kcal, leading to an energy deficit of 7,365 kcal". (Knechtle, 156). To perform well over such a long distance an athlete must take in about 100 calories every 30 minutes. It is possible to do this in liquid form, however the body does not respond well to liquid food while exercising. This is because the body needs to digest the calories slowly to avoid sickness. To go fast, an athlete needs to maximize calorie intake while minimizing stomach sickness.

Finally, there are the transitions. This is where one sport becomes another one. Getting through this phase requires good planning and organization. Having your gear in a place which reduces the amount of unnecessary movement will be a winner. This leads into the second research question, which gender is faster in transition?

Methodology

I will produce some graphs of the data to get a general sense of what types of models would be effective. For example, if the data lends itself to a best fit using linear regression, I will explore that model. If the data is more u-shaped then a binomial distribution would be more appropriate. By minimizing R squared using the method of least squares I can show that the model is correct. Finally, by using a t-test I can do a test of significance. This will give a p-value and can be compared with an alpha level to arrive at a result.

Exploratory Data Analysis

The goal of this analysis was to get a better understanding of the sport of triathlon. Specifically the 70.3 mile distance. I selected this data set because almost all of the data is from this distance. To help answer some of the questions I posed in previous assignments I will present some data visualizations.

Gender: Male or Female athlete AgeGroup: Age group for the athlete for qualifying and placement purposes AgeBand: Age code Country: Country of origin of the athlete CountryISO2: Country code EventYear: Year the event took place EventLocation: Location the event took place SwimTime: Time to complete the 1900 meter swim Transition1Time: Time to complete the transition from swim to bike BikeTime: Time to complete the 56 mile bike Transition2Time: Time to complete the transition from bike to run RunTime: Time to complete the 13.1 mile run FinishTime: Time to complete the entire race

Below is the data I collected.

In [1]:
```
import pandas as pd


df = pd.read_csv("Half_Ironman_df6.csv")
df
```

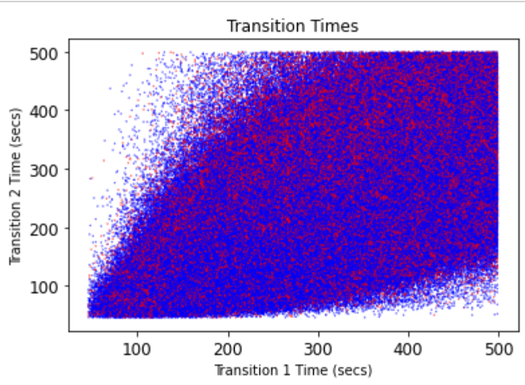
Out[1]:

	Gender	AgeGroup	AgeBand	Country	CountryISO2	EventYear	EventLocation	SwimTime	Transition1Time	BikeTime	Transition2Time	RunTime	Fi
0	M	40-44	40	Andorra	AD	2019	IRONMAN 70.3 South American Championship Bueno...	1679	119	9107	95	5515	
1	M	45-49	45	Andorra	AD	2019	IRONMAN 70.3 South American Championship Bueno...	2070	177	9160	132	6070	
2	M	45-49	45	Andorra	AD	2020	IRONMAN 70.3 Bariloche	1667	161	9891	122	5190	
3	M	45-49	45	Andorra	AD	2019	IRONMAN 70.3 World Championship	1750	183	10363	160	5071	
4	M	40-44	40	Andorra	AD	2019	IRONMAN 70.3 World Championship	2063	182	10065	142	5556	
...	...	...	...	...	...	...	...	...	...	...	...	...	...
840070	M	50-54	50	Zimbabwe	ZW	2015	IRONMAN 70.3 South Africa	2054	261	10527	160	6070	
840071	M	40-44	40	Zimbabwe	ZW	2015	IRONMAN 70.3 South Africa	2449	352	11866	265	8461	
840072	F	30-34	30	Zimbabwe	ZW	2015	IRONMAN 70.3 Steelhead	2171	357	11433	332	7754	
840073	F	35-39	35	Zimbabwe	ZW	2015	IRONMAN 70.3 Budapest	2100	193	10280	233	6148	
840074	F	30-34	30	Zimbabwe	ZW	2015	IRONMAN 70.3 Budapest	2096	244	10630	273	6155	

840075 rows × 13 columns

The scatter plot shows a comparison of the transition times between genders. Blue for men, pink for women. The bottom left has a higher density of blue, implying that men are faster through transition. This could be because men are faster in general and some of the transition is just running.

In [11]:
```
df.plot.scatter(title = 'Transition Times', fontsize=12, colorbar = False, x = 'Transition1Time', y = 'Transition2Time'
, s = .1, c =pd.factorize(df['Gender'])[0], colormap = 'bwr', xlabel ='Transition 1 Time (secs)', ylabel ='Transition 2
Time (secs)');
```



Here I extracted the 3rd fastest finishing time for the 35 to 39 male age group. This is generally the cut off placement for qualifying for the Ironman 70.3 World Championship. For the final paper I will analyze each course specifically, to find a more accurate answer to the question of what time would it take to qualify. Also, it will give me an idea which courses that would have a better chance of qualification.

In [3]:
```
df_M3539 = df[(df['AgeGroup'] == '35-39') & (df['Gender'] == 'M')]
Qualifiers = pd.DataFrame(columns=['EventLocation', 'FinishTime'])
for x in df_M3539['EventLocation'].unique():
    time = df_M3539[df_M3539['EventLocation'] == x].nsmallest(3, 'FinishTime').max().FinishTime
    Qualifiers = pd.concat([pd.DataFrame([[x,time]], columns=Qualifiers.columns), Qualifiers], ignore_index=True)
Qualifiers = Qualifiers.sort_values(by=['FinishTime']).reset_index(drop=True)
print(Qualifiers)
print(Qualifiers.FinishTime.mean())
```

```
                                EventLocation FinishTime
0                        IRONMAN 70.3 Steelhead      14331
1                  IRONMAN 70.3 World Championship      14335
2    IRONMAN 70.3 Middle East Championship Bahrain      14381
3                             IRONMAN 70.3 Dubai      14500
4                          IRONMAN 70.3 Busselton      14506
..                                            ...        ...
190                        IRONMAN 70.3 Chungju      18358
191                           IRONMAN 70.3 Goa      20902
192                         IRONMAN 70.3 Saipan      21436
193                    IRONMAN 70.3 Connecticut      22162
194                           IRONMAN 70.3 Jeju      27995

[195 rows x 2 columns]
15902.174358974358
```

These are the swim, bike and run times as a percentage of the overall finish times averaged together.

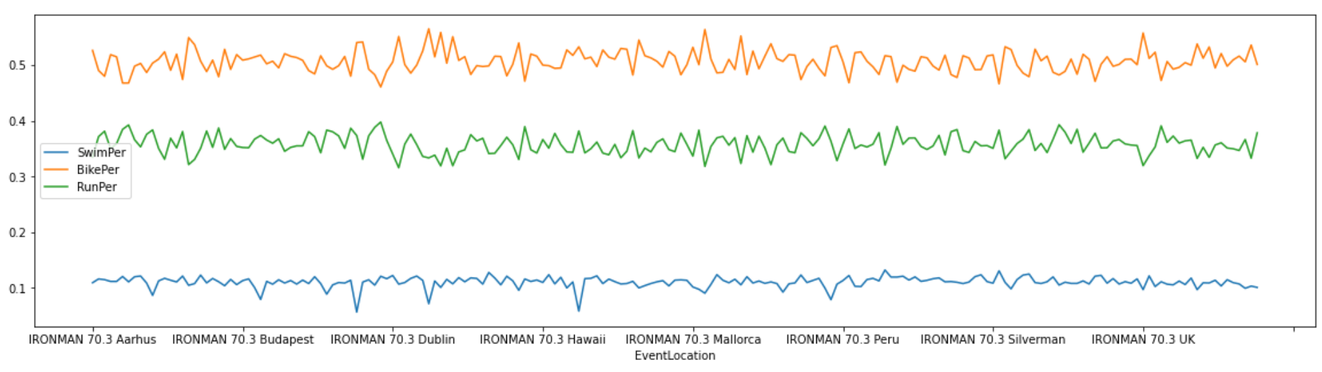
In [4]:
```
df['SwimPer'] = df['SwimTime'] / df['FinishTime']
df['BikePer'] = df['BikeTime'] / df['FinishTime']
df['RunPer'] = df['RunTime'] / df['FinishTime']

print(df['SwimPer'].mean(), df['BikePer'].mean(), df['RunPer'].mean())
```

```
0.11091560725683078 0.5074745411675248 0.3578730663943011
```

In [7]:
```
df.groupby("EventLocation")[['SwimPer', 'BikePer', 'RunPer']].mean().plot(figsize = (20,5))
```

Out[7]:
```
<AxesSubplot:xlabel='EventLocation'>
```



In [ ]:
```

```

Ethical Recommendations

Ethical issues regarding athletes in sports is largely overlooked. The celebrity status of athletes, combined with large salaries, contribute to this problem. However, in the context of professional triathlon, the athletes are neither famous, nor have large salaries. Protecting their privacy and rights should be a main focus. "Bryman (2008) indicates, however, that regardless of the field, discussions concerning ethical issues in social research continue to revolve around the same themes, usefully separated by Diener and Crandall (1978) into four main areas:

- Whether there is harm to participants
- Whether there is a lack of informed consent
- Whether there is an invasion of privacy
- Whether deception is involved" (Harriss, 86).

Most of the information used for sports is widely available on the internet. If the information needed is not easily available, some of it can be gathered using data mining. This creates issues of informed consent. Because most of the information is from official race data, it would be considered public communication, and therefore, not subject to informed consent. However, more data could be collected from other sources and if not cited correctly, could be considered an invasion of privacy as the data collected could be from a personal smartwatch or other biometric sensor device.

The legal ramifications of protecting athletes from already public information is not a simple matter. Classifying Athletes Biometric Data is a common tactic to separate public and private health data. "Currently, no federal laws exist to specifically regulate biometric data collection. Biometric and bio mechanical data are typically not categorized as personal health information (PHI) under existing federal framework, although HIPAA does regulate some biometric data when collected by health care providers. Partially, this gap in the law, particularly with regard to Athletes Biometric Data, is a matter of definitions (of health care purpose, etc.)" (Osborne, 46). This ambiguity means that while the United States government recognizes that there are potential issues with these laws, no one wants to be held accountable for the widespread dissemination of information. Going beyond the official race data seems to be a complicated matter. One that should be handled with care. Using an athletes personal health data to help them improve is one thing. Then using that data later and selling it to corporate entities is another.

Challenges

Filtering the data well enough to be able to accurately model the research question was surprisingly difficult. Pandas data frames do not seem to make it easy to find kth smallest or largest values in a way that can be applied to every row. Losing data seemed guaranteed no matter what I did until I found Pandas.concat and basically rebuilt the data frame in a loop.

Recommendations

A breakdown of the difficulty of racecourses would be good. This could be its' own entire paper with all of the variables involved. Such as weather, temperature, elevation, elevation gain, turns, surface type and size of the transition areas. Gathering this data would involve a lot of work by hand. So I believe only the best, most popular race courses would be used for this analysis at first.

References

Harriss, D. J., and G. Atkinson. "International Journal of Sports Medicine–ethical standards in sport and exercise science research." International Journal of Sports Medicine 30.10 (2009): 701-702.

Osborne, Barbara. "Legal and ethical implications of athletes' biometric data collection in professional sport." MArq. sports L. rev. 28 (2017): 37.

Knechtle, Beat, et al. "What predicts performance in ultra-triathlon races?–a comparison between Ironman distance triathlon and ultra-triathlon." *Open Access Journal of Sports Medicine* (2015): 149-159.

Knechtle, Beat, et al. "Variables that influence Ironman triathlon performance–what changed in the last 35 years?." *Open access journal of sports medicine* (2015): 277-290.

Villiger, E. Ironman 70.3 race data between 2004 and 2020 [Dataset]. https://www.kaggle.com/datasets/aiaiaidavid/ironman-703-race-data-between-2004-and-2020