An Analysis on Men's 100m/200m Butterfly

Authors: Tucker Burhans, Ekta Deshmukh,
Neha Jakkinpali, and Daniel Evora

Introduction:

For our sport, we decided to study swimming. Yes, it is a sport. It's a sport that has been celebrated every four years in the Olympics for decades, but more recently has been gaining attraction from a new professional league. This league is called the International Swimming League (ISL), and it includes 10 teams made up of about 30 swimmers, both men and women, from North America, Europe, and Asia. In a typical swimming match, eight competitors will swim the same distance of the same stroke and race to win, placing first through eighth. The different strokes include butterfly, breaststroke, backstroke, and freestyle. While the distances can be in the form of 100 and 200 meters for butterfly, breaststroke, and backstroke, the distances for freestyle can be 50, 100, 200, and 400 meters. A swimmer can also swim an individual medley, which consists of swimming all four strokes in equal distances in the same race. The individual medley order is butterfly, backstroke, breaststroke, freestyle, and can be swum in the distances of 100, 200, and 400 meters.

Due to the newness of the professional league, and the emergence of more consistent and available swims, our group wanted to take a deeper dive into race execution and what it takes to be the best. We decided to hone in on the men's 100 and 200 meter butterfly in an attempt to gather information about the similarities of a sprint and mid-distance event and how to win a butterfly race. We hoped to determine which variables in a match were key determinants in a swimmer's success, therefore giving us the tools to predict success in an ISL match. In sum, we sought to answer which metrics were related to placing well in a 100/200m men's butterfly race. We used the swimmer's placement in their match to be our measure of success.

First	Last	Date	Match ID	Lap 1 underwater	RT	1st half time percent	Total time	Underwater kicks last lap	Average strokes	Distance	Wingspan	Height	Weight	Placement
Caleb	Dressel	10/16/2020	1	4.9400	0.6400	0.4764	49.6200	6.0000	6.2000	100.0000	193.0000	188.0000	88.0000	2
Chad	Le Clos	10/16/2020	1	4.8200	0.7100	0.4679	49.9700	3.0000	5.8000	100.0000	189.0000	189.0000	84.0000	4
Mateo	Rivolta	10/18/2020	1	4.6800	0.6600	0.4677	50.0100	7.0000	6.6000	100.0000	205.0000	194.0000	68.0000	2
Zach	Harting	10/18/2020	1	4.5000	0.6400	0.4709	51.3300	2.0000	6.6000	100.0000	193.0000	178.0000	75.0000	6
Tom	Shields	10/24/2020	2	4.3400	0.7400	0.4651	49.3000	8.0000	6.4000	100.0000	200.0000	194.5000	94.0000	1

Fifty (50) observations were collected in total. Twenty-five (25) of these observations were of 100m men's butterfly while the other twenty-five (25) were of 200m. Each observation represents one individual athlete's performance in a given race. In each observation, we collected data on twenty-seven (27) unique variables, and we used the following variables in our analysis:

- First - first name of the athlete
- Last - last name of the athlete
- Date - date of the match when the athlete raced
- Match ID - the match number in the ISL 2020 season
- Lap 1 underwater - the time in seconds that the athlete spent underwater after diving off the starting block
- Lap 1 strokes - the number of strokes the athlete took in their first lap after resurfacing from their initial dive

    - Note: this statistic was collected for each lap, however only the strokes on the first lap were analyzed in addition to average strokes
- RT - the reaction time in seconds the athlete took to jump off the blocks after the start of the race
- 1st half time percent - the percentage of time (as a decimal) the athlete spent completing the first half of their race
    - Note: We collected split times for each 50m split (50 and 100 for 100m and 50, 100, 150, and 200 for 200m). We then calculated the proportion of time spent in the first 50m for the 100m race or the first 100m for the 200m race in order to derive the 1st half time percent metric.
- Total time - the total time (seconds) the athlete took to complete the race.
    - It is important to note here that 100m and 200m overall time is included in this variable. Thus, there is a large variability in the data since half the data is from the 100m and half is from the 200m
- Underwater kicks on last lap - the number of underwater kicks the athlete made on their last lap of their race
- Average strokes - the average number of strokes the athlete took over the course of all their laps
    - Note: We collected the number of strokes performed on each lap (4 laps for 100m or 8 laps for 200m) and averaged these records in order to derive the average strokes (per lap) metric
- Distance - this variable kept track of if the observation was for a 100m race or 200m race
- Wingspan - the length in centimeters from one end of the athlete's arms (measured at the fingertips) to the other when raised parallel to the ground at shoulder height
- Height - the length in centimeters of the athlete from their feet to their head
- Weight - the weight in kilograms of the athlete
- Placement - the placement of the athlete in that particular race they took part in (1 for 1st place to 8 for 8th or last place)

The data we collected was derived from three different sources. We collected data such as *first/last* name, *placement*, *RT*, *split* times, and *overall time* from the ISL result sheets. Additionally, we derived the *first half time percent* variable from these variables we collected. However, most of our time in data collection was spent watching recordings of ISL match races which were uploaded to the ISL YouTube account. By viewing these races, we were able to obtain data that led to the following variables: *lap 1 underwater, underwater kicks last lap,* and *lap N strokes*. From *lap N strokes* we derived the *average strokes* variable. Lastly, we obtained physiological statistics such as the *wingspan, height,* and *weight* of each athlete from the ISL's roster.

We made a list of all the swimmers in all the matches for both the 100m and 200m butterfly and randomly selected 25 swimmers (25 observations) for each of the distances while recording the statistics for each swimmer we selected. We were each assigned roughly 12 to 13 swimmers to collect data based on their ISL match videos. Thus, all of our observations and

resulting analyses are only applicable to ISL men's swimming matches with particular emphasis on men's 100 meter and 200 meter butterfly matches. Following the data collection, we each designed a summary statistic table or graph to reveal underlying relationships in our data. From that, we selected the best figures to conduct further analysis.

Summary:

Swimming Statistic	Min	Q1	Median	Q3	Max	Mean	SD
Lap 1 underwater	3.5800	4.4600	4.6750	4.9225	5.5700	4.6622	0.4504
RT	0.5700	0.6325	0.6600	0.7000	0.7600	0.6642	0.0483
50 split	22.8500	23.5575	24.8050	25.5775	26.8400	24.6274	1.0928
100 split	25.4500	26.8225	28.1000	28.9425	30.0000	27.9650	1.1447
150 split	28.1600	28.8800	29.2900	29.5800	31.0900	29.3116	0.6126
200 split	27.7500	29.5200	30.0600	30.1500	31.5500	29.8560	0.8140
1st half time percent	0.4472	0.4654	0.4742	0.4806	0.4852	0.4724	0.0092
Total time	49.1700	50.8125	81.7700	113.5150	118.8400	82.2652	31.7287
Underwater kicks last lap	1.0000	3.0000	4.0000	6.0000	8.0000	4.2200	1.9927
Lap 1 strokes	5.0000	5.0000	5.5000	6.0000	7.0000	5.5200	0.5436
Lap 2 strokes	5.0000	6.0000	6.5000	7.0000	9.0000	6.5200	0.8628
Lap 3 strokes	5.0000	7.0000	7.0000	8.0000	10.0000	7.1400	0.8809
Lap 4 strokes	6.0000	7.0000	8.0000	8.0000	10.0000	7.6800	0.9134
Lap 5 strokes	6.0000	7.0000	8.0000	8.0000	9.0000	7.8400	0.8505
Lap 6 strokes	6.0000	8.0000	8.0000	8.0000	10.0000	8.1600	0.9434
Lap 7 strokes	6.0000	8.0000	9.0000	9.0000	10.0000	8.2800	1.1000
Lap 8 strokes	6.0000	8.0000	9.0000	9.0000	10.0000	8.6800	1.1075
Average strokes	5.2000	6.2000	6.6000	7.1944	7.8889	6.6729	0.7009
Wingspan	176.0000	189.0000	193.5000	200.7500	206.0000	193.7600	8.1455
Height	173.0000	182.2500	187.5000	190.0000	199.0000	186.2100	7.1593
Weight	67.0000	75.0000	83.0000	88.0000	95.0000	81.9400	8.2249
Placement	1.0000	2.2500	4.0000	6.0000	8.0000	4.4000	2.1946

Table 1: Summary statistics of each variable recorded in data collection

Table 1 features summary statistics about each variable we collected data on for our 50 observations. We provided the minimum, first quartile, median, third quartile, maximum, mean, and standard deviation of each variable observed. It's important to note that *total time* has an extremely large standard deviation because it includes the total time variable for both the 100m and 200m races.

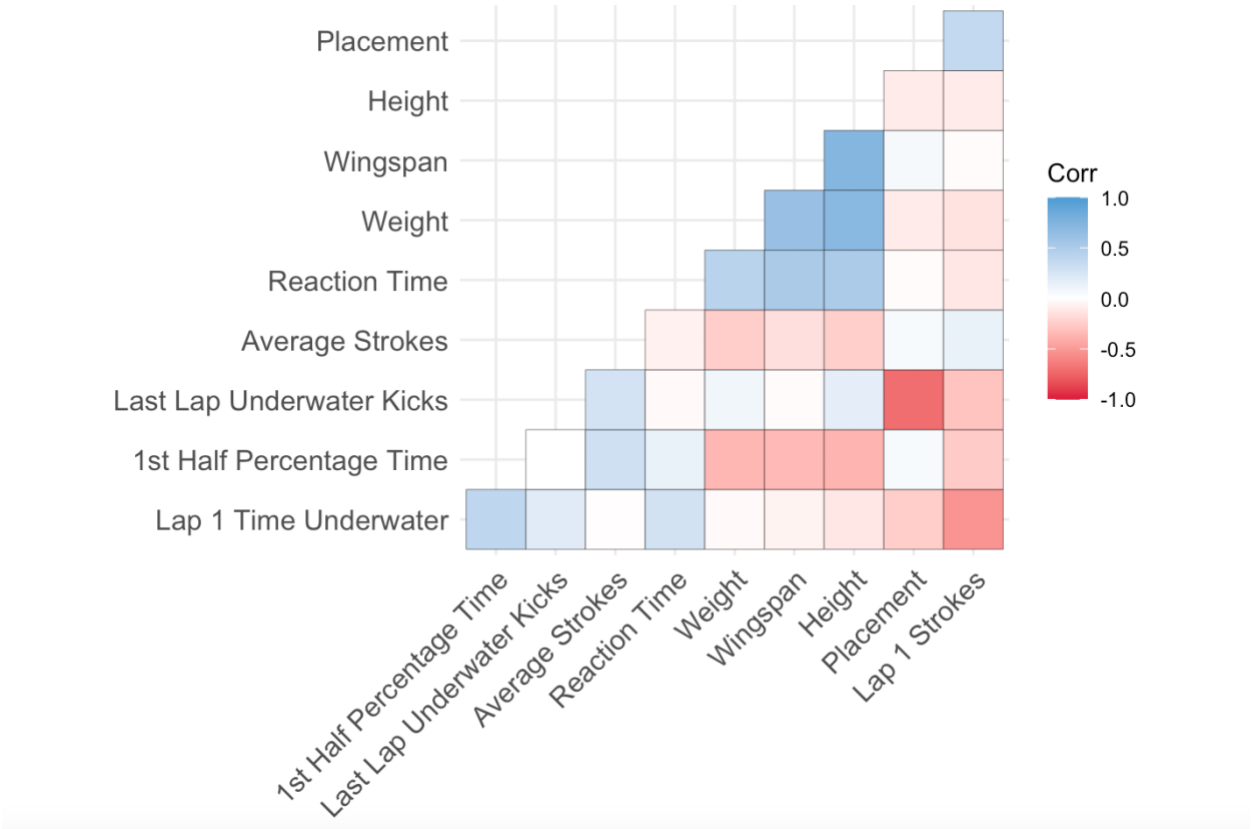

Figure 1: A correlation plot of all variables collected (omitting variables with N/A values)

In **Figure 1**, it is evident that there is a high negative correlation (-0.71) between the number of kicks taken off of the last wall (*last lap underwater kicks*) and the placement in a given race. There is a high positive correlation between wingspan and height (0.73) as well as between weight and height (0.71) of a swimmer.

Since we assumed that tired swimmers would take more strokes to finish (thus having higher average strokes and lower placement), it was unusual to see that the correlation between *average strokes* and *placement* was relatively weak and somewhat positive rather than strongly negative. There was also a low correlation (almost none) between *reaction time* and *placement*, which is likely due to the low variance in the reaction times since all the swimmers are highly skilled and trained to have low reaction times. Ultimately, the *reaction time* was not the most important factor that determined who won. The time a swimmer spent underwater after their initial dive (*lap 1 time underwater*) had a very low correlation (almost none) with the *average strokes* that they took. However, we had hypothesized that the more time a swimmer spent underwater after the first dive, the fewer strokes they would take in their first lap (leading to fewer *average strokes*). This shows that the time a swimmer spent underwater after their initial dive has very little correlation with the average number of strokes, but it does have a negative correlation with the number of strokes they take on their initial lap.

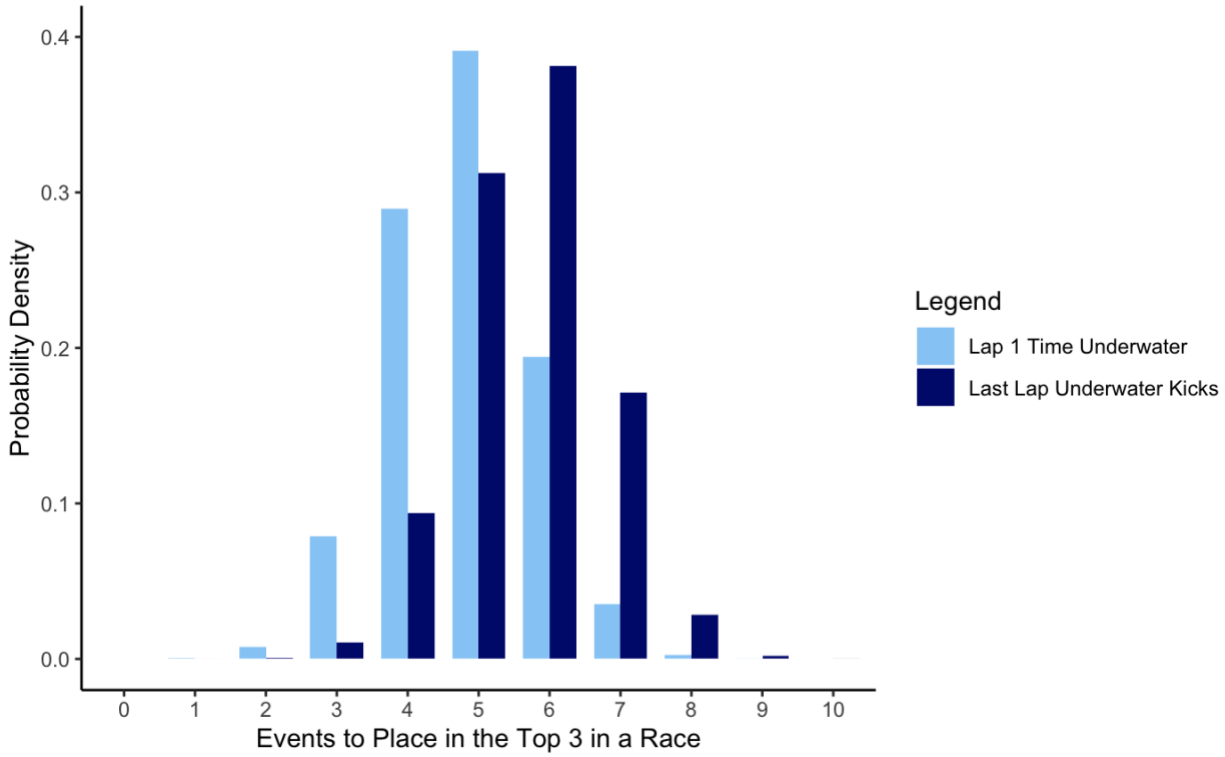

Figure 2: A bar graph of the probability densities that a swimmer will place in the top three of a race based on Lap 1 Time Underwater and Last Lap Underwater Kicks

PDF of a normal distribution: $\frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{1}{2}*(\frac{x-\mu}{\sigma})^2}$

(σ and μ are the variances and means of Lap 1 Time Underwater and Last Lap Underwater Kicks for people that placed in the top three, respectively.)

Since *lap 1 time underwater* and *last lap underwater kicks* are highly correlated to *placement*, we wanted to see what kind of probabilities there were of making the podium (placing top three) if a swimmer did zero through ten kicks as well as stayed underwater for zero to ten seconds. This figure shows how important each second and each underwater kick is in the race, as the probability densities can jump by 0.2 and fall by 0.2 with one underwater kick or one second underwater.

Insights:

From **Table 1**, when viewing other variables besides *total time*, we see very low standard deviations across the board. Contextually, this makes sense. Swimming, especially by event, is very one dimensional. Each athlete is trained to do a few things and to do them exceptionally well. In other sports like basketball, soccer, etc. outcomes are increasingly situational while racing sports have many controlled variables. Furthermore, when looking at *first half time percent*, the standard deviation was 0.0092, meaning each swimmer on average deviated from

the mean of spending 47.24% of the race in the first half by 0.92%. From this, we can conclude that the 100m and 200m races, across both events, are swam very similarly by all racers. This differs from long distance swimming events, where strategic pacing can be an important factor for swimmers who like to keep the tempo fast vs. those who prefer to sit back and then sprint hard to the finish toward the end of the race.

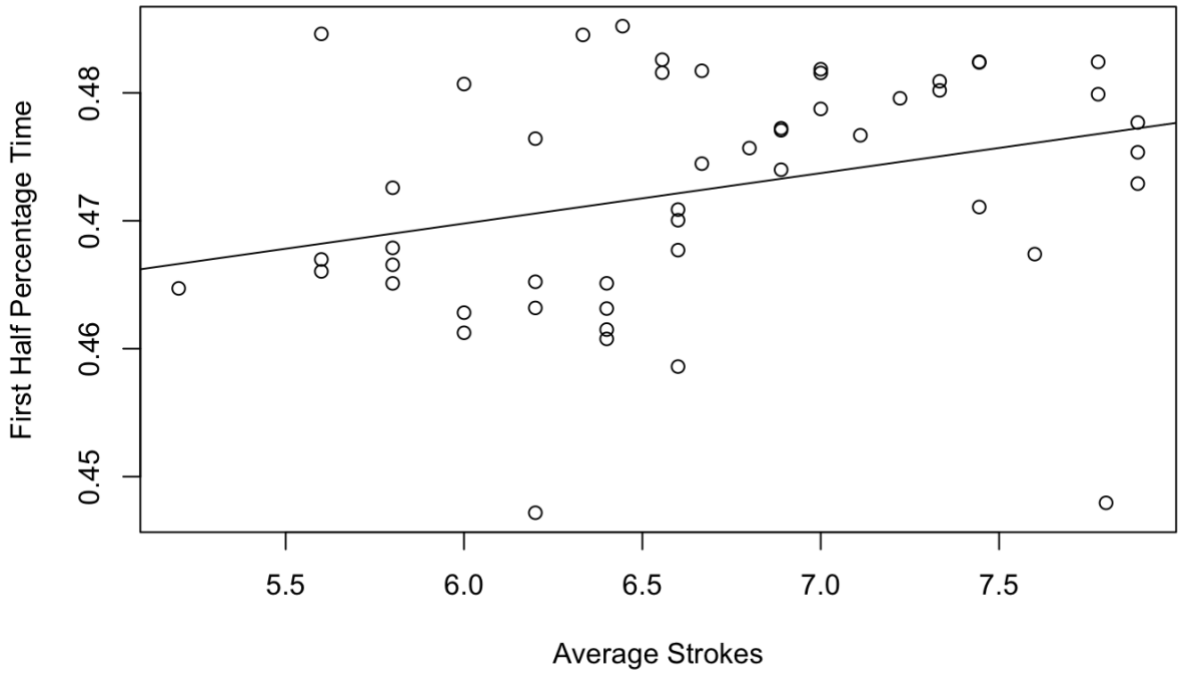

Figure 3: A regression line of average strokes vs. first half time percentage time

Reflecting on **Figures 1 and 3**, we noticed a low correlation between *first half time percent* and *average strokes*. Upon further analysis, the linear model using *average strokes* as the predictor with *first half time percent* as the response variable resulted in this equation:

$$\textit{first half time percent} = 0.4460801 + 0.003951(\textit{average strokes})$$

After performing an ANOVA F-statistic test, we found that the relationship between *average strokes* and *first half time percent* is significant. From the analysis, we can see that the average strokes p-value was 0.0338. The p-value is significant at the 5% level. We set up hypotheses for our test as follows:

H_0: The slope, $ß_1$, relating *average strokes* and *first half time percent* is equal to 0
H_A: The slope, $ß_1$, relating *average strokes* and *first half time percent* is not equal to 0

From the results of our ANOVA F-statistic test, we can reject the null hypothesis. This suggests that the alternative hypothesis is true: *average strokes* is a good predictor for *first half*

time percent even though neither variables are directly correlated. However, it is also important to note how close $ß_1$ is to 0. While this value would suggest no predictive relationship between the two, we also know that that variability in *first half time percent* is also extremely low to 0.

Continuing to consider *first half time percent*, it was thought that the distance of the race would be influential in this statistic. We thought this because a greater proportion of the 100m consists of the initial dive and turnarounds off the wall (which can be the fastest parts of the race). We decided to perform an ANOVA test for means to see if this would hold true with the following hypotheses:

> H_0: The mean *first half time percent* for swimmers in each *distance* (100/200m) are equal
> H_A: The mean *first half time percent* for swimmers in each *distance* (100/200m) are not equal

Our ANOVA test resulted in a p-value of 2.64E-10, which is easily significant at the 5% level. This means that if the mean *first half time percent* in each *distance* for the population was in fact equal, that the probability of us getting these results from our sample was only 2.64E-10. Thus, we can reject the null hypothesis and conclude that the average time spent completing the first half of a race in the 100m is not equal to that spent in the 200m.

Based on **Figure 1**, we noticed relatively low to moderate correlations between *reaction time* and *wingspan*, and *reaction time* and *underwater kicks last lap*. In analyzing it further, the linear model using *wingspan* and *underwater kicks last lap* as the predictors and *reaction time* as the response variable resulted in this equation:

$$\textit{reaction time} = 0.0716376 + 0.0030681(\textit{wingspan}) - 0.0004537(\textit{underwater kicks last lap})$$

After performing an ANOVA F-statistic test, we found that the relationship between *reaction time*, *wingspan*, and *underwater kicks last lap* during the last lap to be significant at the 5% level. From the analysis, we can see that the p-values are very small: for *wingspan*, the p-value was found to be 0.0001399, and the p-value for *underwater kicks last lap* was 0.00345243. We set up hypotheses for our test as follows:

> H_0: The slope, $ß_1$, relating *reaction time*, *wingspan*, and *underwater kicks last lap* is equal to 0
> H_A: The slope, $ß_1$, relating *reaction time*, *wingspan*, and *underwater kicks last lap* is not equal to 0

From our results, we can reject the null hypothesis and have enough evidence to suggest that the alternative hypothesis is true. This states that *wingspan* and *underwater kicks last lap* are good predictors for *reaction time* even though neither variables are directly correlated. The VIF

values for both *wingspan* and *underwater kicks last lap* are low. This shows that the correlation between the predictors is slim and therefore multicollinearity is less of an issue.

This proves to be valuable information, as every swimmer is looking to minimize their reaction time through ways that can be improved upon. Since wingspan cannot be changed, this points to the importance of underwater kicks and how it can impact multiple variables within a race.

Based on **Figure 2**, we wanted to take a deeper look into the relationships between *placement* and the *kicks taken in the last lap* and the *time spent underwater in the first lap*. We wanted to know if these variables were reliable predictors for *placement*, and what the positives and negatives there are of taking more or less *kicks underwater in the last lap* or *time underwater in the first lap*. After further analysis, we came up with a linear model that gave us the best equation to predict placement:

placement = 10.2533 - 0.7586(*underwater kicks last lap*) - 0.5689(*lap 1 underwater time*)

After performing an ANOVA F-statistic test, we found that the relationship between *underwater kicks last lap*, *lap 1 underwater time*, and *placement* to be significant. The p-values are both significant at the 5% level as the values are both less than 0.05. The p-value of *underwater kicks* in the last lap is 7.378e-09 while the p-value of lap *1 underwater time* is 0.00764. We set up hypotheses for our test as follows:

H_0: The slope, β_1, relating *underwater kicks last lap*, *lap 1 underwater time*, and *placement* is equal to 0
H_A: The slope, β_1, relating *underwater kicks last lap*, *lap 1 underwater time*, and *placement* is not equal to 0

From the results of our ANOVA F-statistic test, we can reject the null hypothesis. This suggests that the alternative hypothesis is true that *underwater kicks in the last lap* and *lap 1 underwater time* are good predictors for placement even though neither variables are directly correlated.

Reflection:

Looking back on the work we've completed, there are several elements that hindered our ability to find more interesting relationships between 100/200m butterfly race statistics and success. The most obvious is the small sample size and low variability, with statistics which are so close numerically and only 50 observations, it was definitely difficult to find race factors that separated the most elite swimmers. Another improvement that can be made on our research is the potential for human error. Variables such as *lap N strokes* and *lap 1 underwater* time were recorded watching race film and phone stopwatches. When camera angles were suboptimal, recording these variables was very challenging. Lastly, we would have liked to record the total amount of time spent underwater in each race. Kicking off of the wall is oftentimes the fastest

portion of these events and we believe recording this as a variable would have led to interesting findings in our data. However, it is important to recognize that this too would have increased the potential for human error.

As a result of our analysis, we can draw an interesting conclusion from the information we collected holistically: athletes in the 100/200m men's butterfly swimmers in the ISL swim these events very similarly to one another. This can be attributed to the nature of the sporting event, as it is a sprint event and there are not many differing approaches to how to win besides swimming as fast as you can for the entire duration of the race. Analogous to the 100/200m events in track racing, there really isn't too much that a swimmer can do besides go as fast as they can. However, it is important to note that there are other important factors in track that don't convert to swimming such as cadence and stroke length.

An interesting fact about our data we discovered was that wingspan and underwater kicks in the last lap are good predictors for reaction time. However, intuitively there doesn't seem to be a reason, and this relationship we found could be attributed to the small sample size.

Additionally, throughout the analysis we saw an extremely low variability in our data and we believe the reason for this is that we are dealing with a select group of professional athletes who are excellent at a very particular race. While many of our coefficients were close to 0, it is important to keep in mind the low variability of the data and that the differences in placements, reaction times, or first half time percent between swimmers came down to a matter of milliseconds. Therefore, these coefficients for our predictors do provide insights for close races such as those in the ISL. In other words, these athletes are all at the upper threshold for performance for these events, and their variability is likely to be low, resulting in lower coefficients. In future research, it would be interesting to look at athletes in these events at different skill levels in order to increase variability and see the key differences that separate average swimmers from the professionals.