Title
Statistical Analysis of The 2016 and 2017 NCAA Division-I Swimming Championships

Author
Kaunitz, Sarah Louise

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Statistical Analysis of
The 2016 and 2017 NCAA Division-I Swimming Championships

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Sarah Louise Kaunitz

2020

ABSTRACT OF THE THESIS

Statistical Analysis of
The 2016 and 2017 NCAA Division-I Swimming Championships

by

Sarah Louise Kaunitz
Master of Applied Statistics
University of California, Los Angeles, 2020
Professor Frederic Paik Schoenberg, Chair

This paper applies the implementation of web-scraping to create a single new dataset composed of eight separate competition results datasets. Exploratory analysis will be performed in large to identify the measurable reasons why most swimmers perform worse at the fastest collegiate competition in the nation. Additionally, using forward and backward stepwise variable selection, the impact of various factors on the outcome variable time difference will be studied. Machine learning algorithms such as ridge regression and lasso method will create models that predict the time difference between entry time and final time of a swimmer's race. The mean squared error value will evaluate the overall performance of the models. Although many variables are created and used to best fit the final ridge regression model, there are unmeasurable factors that must be taken into account to accurately describe what impacts how fast a swimmer goes at the competition.

The thesis of Sarah Louise Kaunitz is approved.

Yingnian Wu

Robert Gould

Frederic Paik Schoenberg, Committee Chair

University of California, Los Angeles

2020

To my family . . .

who has always supported me in all facets of life

and pushed me to be the best

student, athlete, and person I can be

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

(Acknowledgments omitted for brevity.)

CHAPTER 1

Introduction

1.1 Background

How to be faster in swimming is simple, just "drop time". Decrease the number of seconds the scoreboard displays in fluorescent neon lights next to your name. This simplistic statement, however, is much easier said than done. Swimmers train year-round for a chance to compete at the highest level competition in collegiate swimming; the NCAA Division-I Championships. With a staggering 9,609 total NCAA Division-I swimmers in the 2019-20 season spread over 196 schools across the country [1], having the opportunity to be selected as one of the few elite swimmers in the country is an honor in itself. To qualify, athletes can only get official invites based on their individual swims. In essence, the qualification process roughly selects the top 38 women and 29 men in each event will get an invite [2] based on the race times they display earlier throughout the season. These displayed times can be swum at any point throughout the season, and will be considered in this paper as a swimmers entry time. As the selection process follows, the NCAA invites the same number of overall swimmers every year with 270 men and 322 women in attendance. Depending on how many of those 270/322 athletes qualify in multiple events, the numbers can range as to how many entries in each event get invited [2]. This typically allows between 35-42 swimmers to be entered in each event. As a result, the number of heats for each event will differ because each heat will only contain eight swimmers, and each swimmer will only swim once in their assigned heat and lane. There are a total of eight lanes to race in, eight swimmers in each heat, and the heats are filled from the last one to the first one. This means that all heats

will have exactly eight swimmers in each one, except for the very first heat, which may have less people if the number of heats is not evenly divisible by eight.

The NCAA championship is the most competitive competition in all of college swimming and takes place once a year, during the postseason in March. Once selected to attend the competition, many factors may impact how a swimmer will perform once they compete. The idea of putting the best athletes all together in one place, at one competition seems like an electric experience with no shortcomings of fast swimming. Little research has been conducted on the sport of swimming at this level given these variables thus far. As a swimmer myself, and having had the honor to be selected to attend, I have firsthand experience with the training, atmosphere, and results produced at the competition. The goal through this paper is to uncover the underlying reasons behind why the final swimming results do not accurately reflect the fast entry times at the NCAA competition.

The entire sport of swimming revolves around time. How quickly can you jump off the starting blocks, how fast can you kick underwater, and who can get their hand to the touchpad first. Every swimmer at the NCAA competition is of the highest caliber, and examining the range of entry times between the first and last place swimmers can be a mere few tenths of a second. However, even with relatively similar entry times, the drastic difference in final times creates an unexplored question to be answered. Is it true that the fastest people in the country swim slower at the NCAA competition? This paper will discuss several factors of the NCAA competition and events raced by these talented athletes. More specifically, the focus will lie on two polar opposite events – the 50 yard freestyle and the 1650 yard freestyle, for both men and women. The data collected will represent the results from two different years, 2016 and 2017.

The objective of this research paper is to discover insights and find the best predictive model that can identify the most important variables that impact the time difference between entry and final time at the competition. As a result, both swimmers and coaches will be able to harness the information provided to advantageously plan for a positive outcome at the

meet by strategically training before the competition. To properly achieve the paper goals, the following content can be separated into three major categories:

1: Dataset creation

2: Exploratory data analysis

3: Predictive modeling

1.2 Data Collection and Creation

Due to it's more niche following, the swimming world is often not present in the limelight for statistical research or analysis. This leads to a lack of datasets available and in order to create a well-executed study focused on the NCAA competition, a proper dataset must be created. The first two steps for collecting and creating the dataset are to find the competition results online, and then web-scrape the viable information needed. I chose to examine eight datasets in total:

· 2016 NCAA Men's 50 yard freestyle

· 2016 NCAA Men's 1650 yard freestyle

· 2016 NCAA Women's 50 yard freestyle

· 2016 NCAA Women's 1650 yard freestyle

· 2017 NCAA Men's 50 yard freestyle

· 2017 NCAA Men's 1650 yard freestyle

· 2017 NCAA Women's 50 yard freestyle

· 2017 NCAA Women's 1650 yard freestyle

These two specific events were chosen because they are the shortest and longest events possible to race, and it will be interesting to examine the debate between sprint and distance events as it relates to people swimming faster.

Each results page had minor differences that required individualized approaches in the creation of a dataset, however, shared practices among all events follow:

1: Install readr, readxl, xml2, rvest, and stringr R library packages.

2: Use the appropriate libraries and commands to import the data from each website.

3: Unpack and filter the data to concatenate the information on the variables available into eight separate datasets.

4: Variable manipulation and refactoring to achieve the correct data type. Change the date to its appropriate type either strings, characters, numeric, and factors. Use filter, unlist, apply, substr, gsub, and str_split functions to extract the correct variable information.

5: Concatenate data from multiple sources on the same event.

6: Create additional variables including time_diff, finals, heat_place, eventtype, droptime, entry_heat_place, heat_place_movement, lane_loc, and std_time.

7: Combine all eight datasets into one to begin analysis.

1.3 Data Manipulation

To properly construct my dataset, the first step is to unpack the available results information from each webpage. Each of the eight datasets contained the year of competition, the gender of the swimmers, the final place the swimmers ranked in each event, the swimmer's name, their year in school, the school they attended, their entry time, and final time. In addition to the main data that was imported, I was very interested in the heat, lane, entry heat place, and final heat place of each swimmer for each event. I hard-coded these into an excel file and then imported that to be combined with the existing data. Once all the basic data

had been imported, additional variables were created including the time difference between a swimmers entry time and final time, whether or not they made finals based on the place they ranked after the race, their place movement after the race, the lane location being either within the inside four lanes or the outside four lanes, whether or not their final time was faster than their entry time denoted as "dropping time" in the swimming world, and finally, I standardized the time difference as a z-score because of the race time between the two events I compared. After all of the data scraping, manipulation, and variable creation, the final dataset included twenty variables. The final step of the data cleaning and manipulation included creating mean values of the numeric variables- std_time (standardized time difference), heat_place, and entry_heat_place. There was no missing data to deal with and a total of 365 rows were collected.

Subsets of data: To properly work with and interestingly analyze the data, smaller subsets of data from the full dataset were created. These subsets are used to control for certain variables, including competition year, gender, swimmer's year in school, and event distance. Of the 193 schools with swimming programs, only 64 schools had swimmers qualify for the competition. To control for the fact that some schools having a large number of swimmers and others having only one or two swimmers invited, an additional subset of data was extracted to include who schools who qualified more than twenty swimmers, which led me to create a dataset of the top three schools.

1.4 Data Codebook

The dimensions of the final data set are 365 total swim races (rows), by 20 variables (columns). There is no missing data as none of the swimmers missed their races. The variables are split between either numeric or factor data type. Of the twenty variables, eight variables are numeric, twelve are factors, and of those twelve, six are binary. The numeric ones are also split between either being full integers or just numeric because of the decimal included in the race time. The most important variables that will be studied are 'class', 'entry_time', 'final_time', 'event', 'year', 'heat', 'lane', 'heat_place', 'gender', 'droptime', 'finals', 'heat_place_movement', 'lane_loc', and 'std_time'. These variables will be examined heavily during the exploratory analysis section and will be crucial in creating valid models to predict the outcome variable of 'std_time'.

1.4.1 Lane and Heat Assignment Process

The manner in which swimmers are placed into their assigned heat and lane for each race is an intricate system designed to make the playing field even. The same technique is followed throughout all swimming competitions and considered a standard practice. As discussed above, the number of heats will vary for each event because it is dependent on the number of swimmers entered in that event. The number of lanes in each heat will never change, with the exception of the first heat that may have fewer than eight swimmers racing in it. There is never more than eight swimmers per heat, and each swimmer competes only once in their assigned heat and lane. The manner in which the heats are filled is from fastest to slowest. The very last heat is filled first with the top eight swimmers, the second to last heat is with swimmers ranking nine through sixteeen, and so on, until the first heat has the leftover amount of swimmers in it, up to eigth. Within each heat, swimmers are positioned from fastest to slowest from the inside lanes moving outwards. The fastest swimmer in each respective heat will be in lane 4, the second fastest will be placed in lane 5, the third will be

in lane 3, followed by lane 6, lane 2, lane 7, lane 1, and the eighth person in each heat will be positioned in lane 8. This is repeated for each heat. The fastest entered swimmer will be positioned in lane 4 of the last heat, and the slowest swimmer will be in the outermost lane in heat 1. This is why the entry time is so important when entering a competition. The entry time data a swimmer is associated with determines their entry heat and lane position for the race.

Variable	Structure	Description
rank	Numeric - Integer	Rank of the swimmer in their respective event based on final_time displayed
name	Factor w/ 310 levels	Name of the swimmer
class	Factor w/ 4 levels	Year in school
school	Factor w/ 64 levels	College the swimmer attends
entry_time	Numeric	The time the swimmers entered the competition with
final_time	Numeric	Final time swum at the competition
event	Factor w/ 4 levels	Swimming event name including gender
year	Factor w/ 2 levels	Year of competition
time_diff	Numeric	Variable created by subtracting entry_time from final_time. Is the time difference between final time a swimmer displays at the meet and entry time the swimmer entered the meet with in seconds
heat	Factor w/ 8 levels	Heat number of event that swimmers race in

Variable	Structure	Description
lane	Factor w/ 8 levels	Lane number the swimmer raced in
heat_place	Numeric - Integer	The numeric rank the swimmer placed within their heat
gender	Factor w/ 2 levels	Gender of the swimmer
eventtype	Factor w/ 2 levels	Distance of the event
droptime	Factor w/ 2 levels	Whether or not a swimmer swam faster and dropped time from entry time. 0 = No (swam slower), 1 = Yes (swam faster)
entry_heat_place	Numeric - Integer	The rank the swimmer was placed with their entry time into each heat
finals	Factor w/ 2 levels	Whether or not the swimmer made it into the finals round. Only 16 spots are available for each event in the finals round
heat_place_movement	Numeric - Integer	The rank movement of a swimmer within their heat after racing. Tracks the numeric value of if a swimmer ranked higher or lower than the place they were entered as
lane_loc	Factor w/ 2 levels	Lane location split between the four outside lanes (1,2,7,8) and the four inside lanes (3,4,5,6)
std_time	Numeric	Standardized value of time difference (time_diff) variable between final_time and entry_time

CHAPTER 2

Exploratory Data Analysis

Since this data is new and being analyzed for the first time, the exploratory data analysis constructs the majority of this paper. The data has a strong foundational element of being very normal when speaking in terms of distribution. I know before doing any type of analysis that for the variables including heat, heat place, lane, number of rows for each event, finals and lane locations, that because of the unique NCAA selection process and the limited number of heats and lanes, that there won't be any major issues in these regards.

Initial findings in the project produce not so seemingly positive results. From table 2.1 below, out of 365 total swims recorded, in more than half of those races, swimmers added time, meaning their final_time was slower than their entry_time.

All swimmers and time difference		
-1 (drop time)	0 (same time)	1 (add time)
125 (34.25%)	12 (3.28%)	228 (62.47%)

Table 2.1: Count and Percentage of swimmer who Drop Time, Stay the Same, and Add Time.

Table 2.2 displays the number of people who placed in ranks 1 through 8 in their respective heat dependent on the lane they swam in. The red highlighted portions from the table indicate what place most people ranked in each lane. The codebook and description in 1.4.1 describe what the expected lane should be for each place based on the swimmers' entry time. For first and second place within each heat, the initially expected lanes to place first and

second do in fact follow the entry place with the majority of heat winners to come from lane 4, and the majority of second place finishers to come from lane 5. Interestingly enough, for third place, we see rather than lane 3 being the majority of that place, lane 6 had the most amount of third places by almost double all the other lanes. For fourth place, we see a split between lane 6 and lane 3, fifth and sixth both do not conform to the entry heat place standard, seventh place is split between lane 1 and lane 2, and lane 8 has the most people ranking eighth in their heat, which is expected.

COUNT of heat_ heat_place									
lane	1	2	3	4	5	6	7	8	Grand Total
1	2	2	4	5	5	9	9	5	41
2	7	6	4	6	8	5	9	3	48
3	5	6	6	8	9	6	7	3	50
4	18	8	4	2	4	5	1	4	46
5	12	16	5	7	4	2	1	1	48
6	1	7	15	8	6	2	3	5	47
7	4	6	8	7	6	7	4	4	46
8	2		3	4	5	6	8	11	39
Grand Total	**51**	**51**	**49**	**47**	**47**	**42**	**42**	**36**	**365**

Table 2.2: Count of Swimmers for each Heat Place in each Lane (1-8)

Figures 2.1 and 2.2 contain information about five different variables. The y-axis is the lane_loc variable where "0" denotes the outside lanes, and "1" denotes the inside lanes. The plots are also split by event including M 1650 free, M 50 free, W 1650 free, and W 50 free. The x-axis denotes the class/year in school of the swimmers going from freshman to senior, left to right.

We have a total of 32 categories based on the categorical combinations possible. We see in the men's 1650 yard race, for outside lanes, seniors rank as the best class, however, for inside lanes, sophomores rank better. Another really interesting takeaway is seen in both of the women's races; inside lanes consistently place better than the outside lanes which are expected, but in both of the men's races, this trend is not as apparent. Looking more in-depth at just the average standardized time difference, we see two bright yellow balloons in the women's 50 yard race for the inside lane sophomores and an even brighter yellow dot for the inside lane freshmen in the men's 50 yard race meaning they added the most time

on average.

In figure 2.1 the size of the balloon is the average entry heat place for each combination of swimmers grouped by class, lane_loc, and event. The color of the balloon is the average entry heat place for each category of class, lane_loc, and event. A bigger balloon means that that group of swimmers came in ranked lower compared to swimmers with smaller balloons. A similar idea follows for the color of the balloon. If a balloon is brighter yellow, that means that the average standardized time is more positive and that group of swimmers on average added time and went slower than their initial entry time. And the reverse is true for darker purple colored balloons.

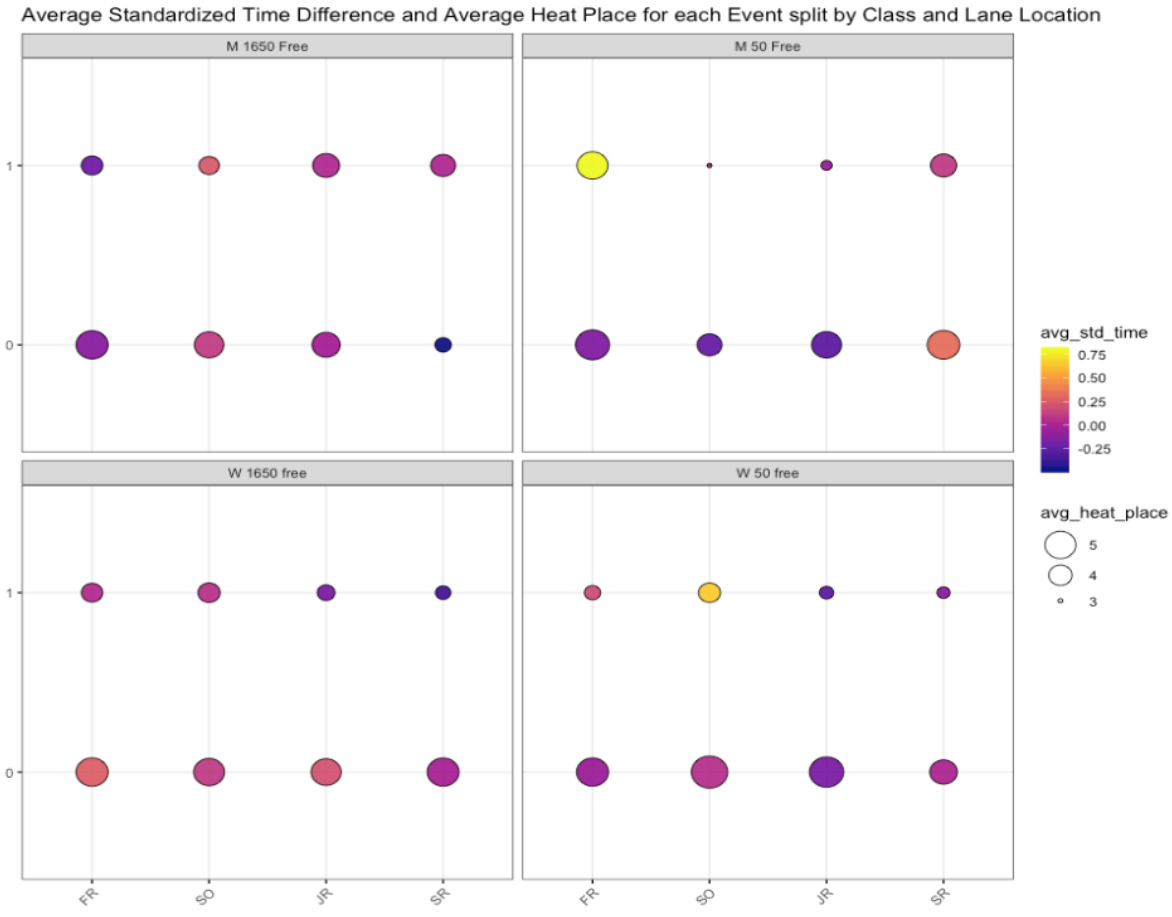

Figure 2.1: Average Standardized Time Difference and Average Heat Place for each Event split by Class and Lane Location (0 = Outside four lanes, 1 = Inside four lanes)

In figure 2.2 the color of the balloon is the average standardized time difference between final_time and entry_time for each category of lane_loc, class, and event. The more yellow a balloon is the higher the standardized time difference that category has on average. The more dark purple denotes a lower average standardized time difference. Looking at the size of the balloon, the different size comes from the difference in an average heat place for each category. The bigger the balloon, the slower and lower-ranked, are the swimmers in that class on average for each category. Ideally, a swimmer wants to have a smaller value for time, which means a higher overall rank.

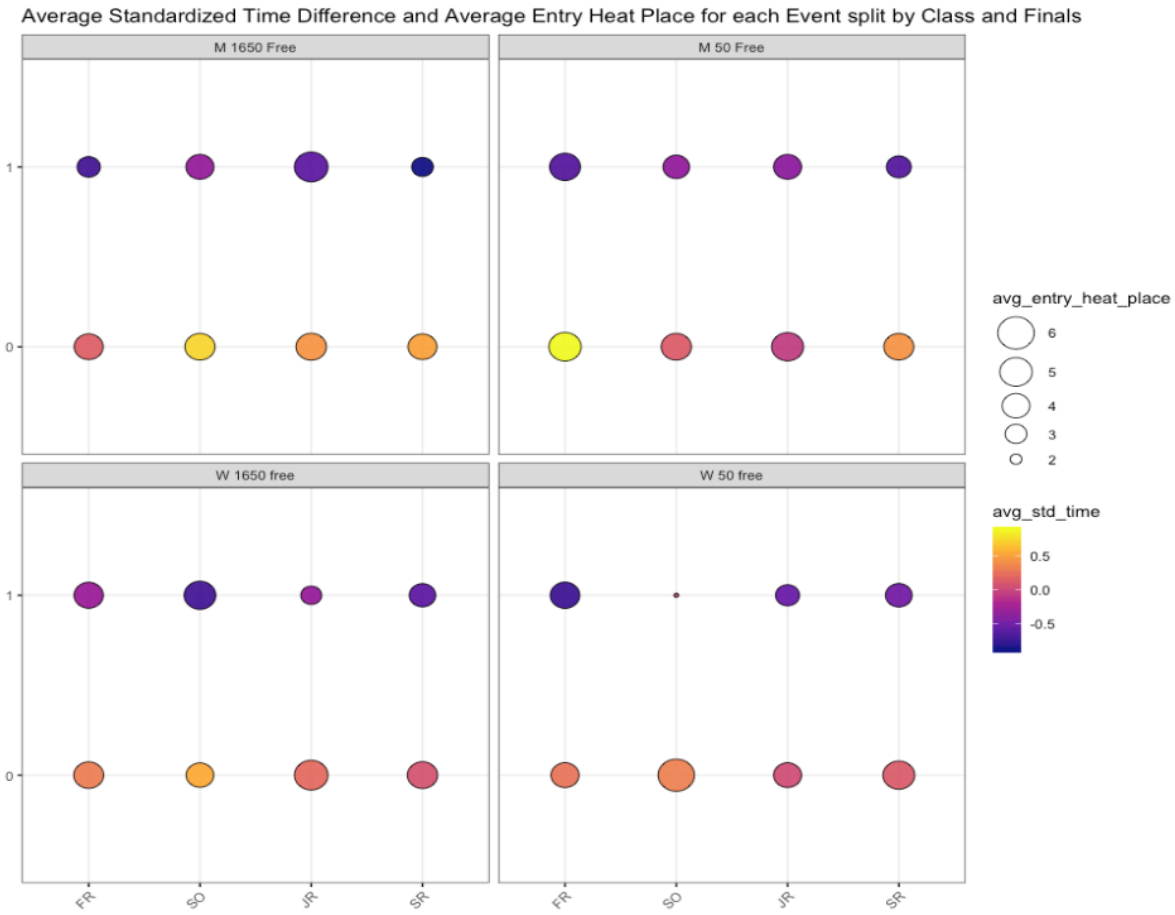

Figure 2.2: Average Entry Heat Place and Average Standardized Time Difference for each Event split by Class and if the swimmer made it in Finals (0 = No, 1 = Yes)

The outcome we are interested in is the time difference between entry time and final

time. This outcome variable of the time difference for each event is the result of entry time subtracted from the final time. Because of the dependent nature, the outcome variable has on two values that are being measured on the same person, I decided to perform a paired t-test on the variables.

T-Test is used to determine how significant the difference between the two groups are, and if those recorded differences (measured in means) could have happened by chance [3]. The null hypothesis of this data is that the mean time difference between the paired variables, entry_time and final_time, is equal to zero. The alternative hypothesis is that the mean time difference is not equal to zero. The larger the t-score, the more difference there is between groups, and the more likely it is that the results are repeatable. The smaller the t-score, the more similarity there is between groups [3]. The way to tell if the resulting t-score is small or big is by looking at the associated p-value that goes with it.

P-Value is the probability that if the results from the data occurred by chance, the outcomes as extreme or more extreme than the observed data would happen. P-values assume that chance is the only factor at play. These values can be read as a percentage between 0 to 100 and are expressed in decimal format as part of the t-test outcome. The lower the p-value, the lower the probability of obtaining a result like the one that was observed if the null hypothesis was true. Thus, a low p-value indicates decreased support for the null hypothesis [4]. The threshold used to determine whether a p-value is small or not is up to each person and their experiment, however, common practice chooses 0.05 as the value to accept data as not being produced by chance. If the p-value is less than significance level $\alpha = 0.05$ ... we can reject the null hypothesis and conclude that entry_time is significantly different than final_time.

The assumption of a t-test must be met before proceeding. The paired sample t-test has five main assumptions:

· The dependent variable must be continuous.

· The pairs of observations are independent of each other. The observations within each pair are dependent.

· The dependent variable should be approximately normally distributed.

· The dependent variable should not have any outliers [3].

· The sample be randomly selected from a larger population.

The data meets the first four assumptions. In regards to the last assumption, this data does not meet the criteria due to the two selected events nature of the project, however, an "as-if" analysis is still valid; stating if assumed for the data to be randomly sampled, then the assumption would be met and a t-test is a valid option. The histogram in figure 2.3 displays the distribution of the data confirming it to be normally distributed.

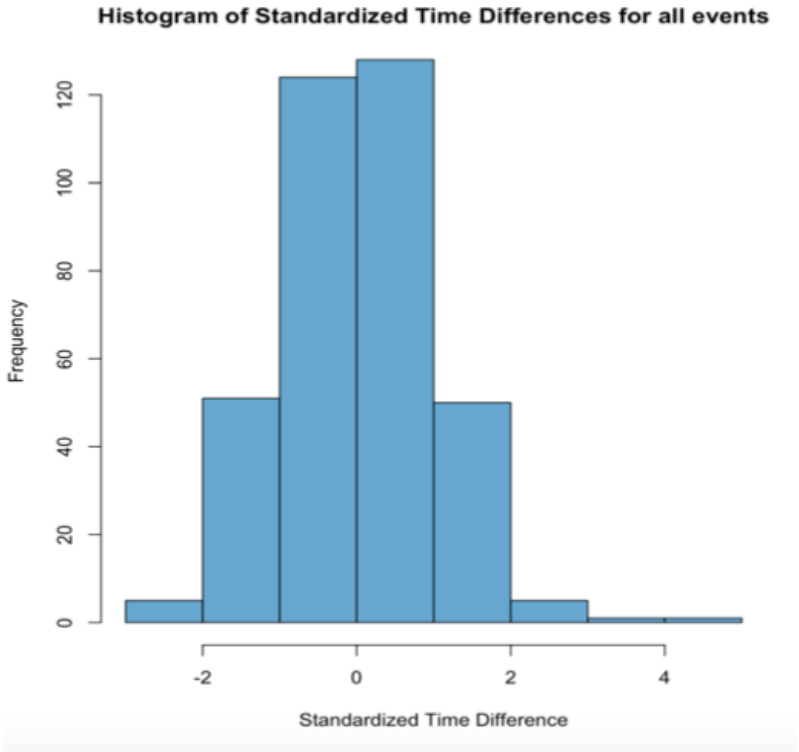

Figure 2.3: Histogram of Standardized Time Differences for all Events

The negative t-score outputs of the t-test are displayed in table 2.3. After viewing all of the t-test statistics, a conclusion can be stated that due to the short nature in terms of time in seconds of the 50-yard race, every tenth and even hundredth of a second is important and the significance can change much more rapidly as compared to the long-distance 1650 yard race. Additionally, three of the four events had significant time differences between entry time and final time as an outcome of the significance $\alpha = 0.05$ p-value analysis. All four events had a negative mean of difference, meaning that on average of all the swimmers in those events there was faster swimming at the competition than the entry times.

Event	T-Score	P-value	Degrees of Freedom	95% Conf Int.	Mean of Differences
W 1650 free	-2.7978	0.006334	87	[-3.945, -0.066]	-2.306
W 50 free	-8.6323	6.22E-14	107	[-0.189, -0.118]	-0.154
M 1650 free	-1.288	0.202	70	[-4.163, 0.896]	-1.633
M 50 free	-2.6219	0.01015	97	[-0.103, -0.014]	-0.058

Table 2.3: T-Test Outcome

In figure 2.4, when comparing class or year in school of a swimmer at the NCAA competition, an initial guess is for the freshmen to add the most amount of time as a result of their first experience attending this meet. The NCAA championship meet is the highest-level competition for collegiate swimming and the inexperience of attending a high-pressure meet could account for time added by freshmen. At the same time, being chosen to swim in the meet is very difficult, so many people do not have the opportunity to attend until their sophomore, junior, or even senior year in college. The plot demonstrates that each respective event has the same effect on both juniors and seniors. In the 1650 freestyle, the swimmers drop time as they mature, and in the 50 freestyle, both genders of swimmers add time as they mature.

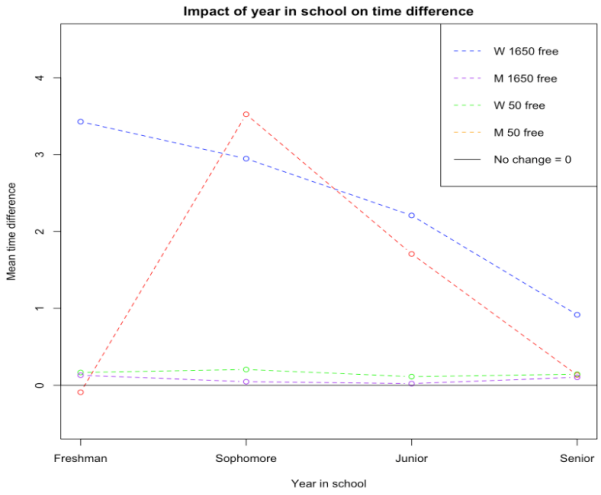

Figure 2.4: Plot of the Mean Time Difference for all Classes in each Event

To examine the amount and percentages of people who added, stayed the same, and dropped time in each event, a simple yet effective test is through a mosaicplot in figure 2.5 and proportion table 2.4. The proportion tables look at the number of people dropping and adding time in each event. Table 2.4 shows that the women's 50 free in 2017 has the highest proportion of people adding time with nearly 41.67% , and only 5.56% dropping while 1.85% stayed the same. In contrast, the men's 1650 free in 2017 had the highest proportion of time drops with 28.17% , while 21.13% added time and only 2.82% stayed the same. A mosaicplot plot was created to help illustrate the results, which displays what percentage of swimmers within each event added, stayed the same, or dropped time for each year. In examining the mosaicplot a common trend is found, being that the women's 50 free was the event that had the most time added for both the years 2016 and 2017. Additionally, the men's 1650 free had the least amount of time added and most amount of time dropped.

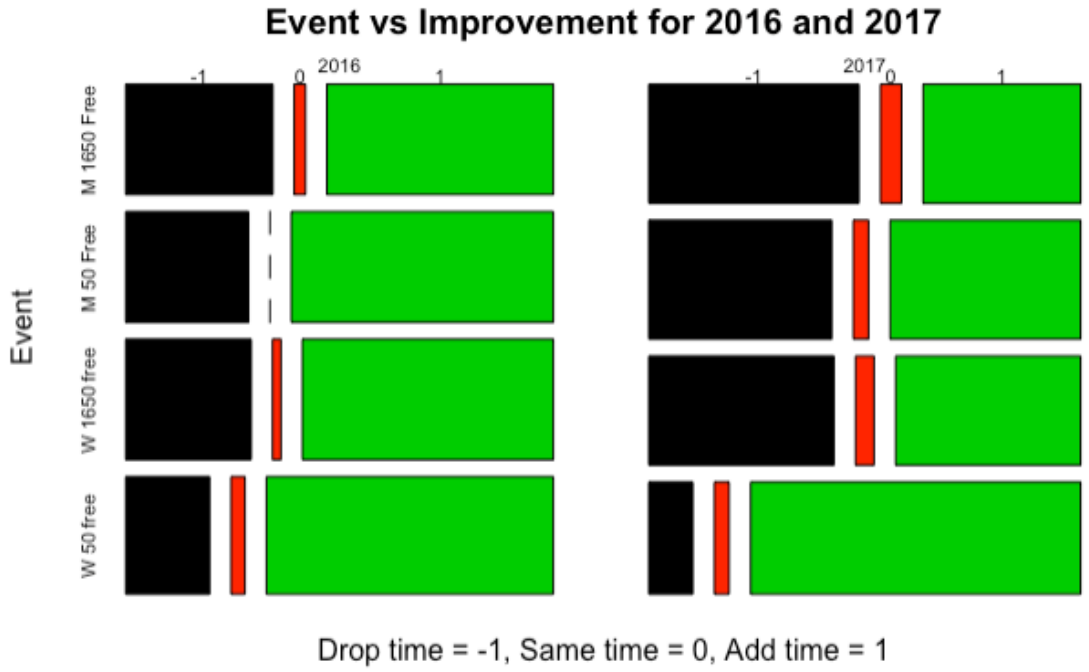

Figure 2.5: Mosaic Plot: Distribution of Drop time, Same time, Add time for each Event and Year

Drop Time = -1	**M 1650 Free**	**M 50 Free**	**W 1650 free**	**W 50 free**
2016	18.31%	15.31%	17.05%	11.11%
2017	28.17%	24.49%	22.73%	5.56%
Same time = 0	**M 1650 Free**	**M 50 Free**	**W 1650 free**	**W 50 free**
2016	1.41%	0.00%	1.14%	1.85%
2017	2.82%	2.04%	2.27%	1.85%
Add Time = 1	**M 1650 Free**	**M 50 Free**	**W 1650 free**	**W 50 free**
2016	28.17%	32.65%	34.09%	37.96%
2017	21.13%	25.51%	22.73%	41.67%

Table 2.4: Values corresponding to Mosaic Plot sections

In figure 2.6 representing the 50 yard race, there was an overwhelming amount of people who added time. This supports the initial belief that if a race is shorter, more people add time. Also, an interesting trend has been revealed; by moving up a lane starting from 2 and ending at 7, the percentage of people who drop time is clear. Regardless of the lane number a swimmer is placed in, a majority of swimmers added time across all possible lanes.

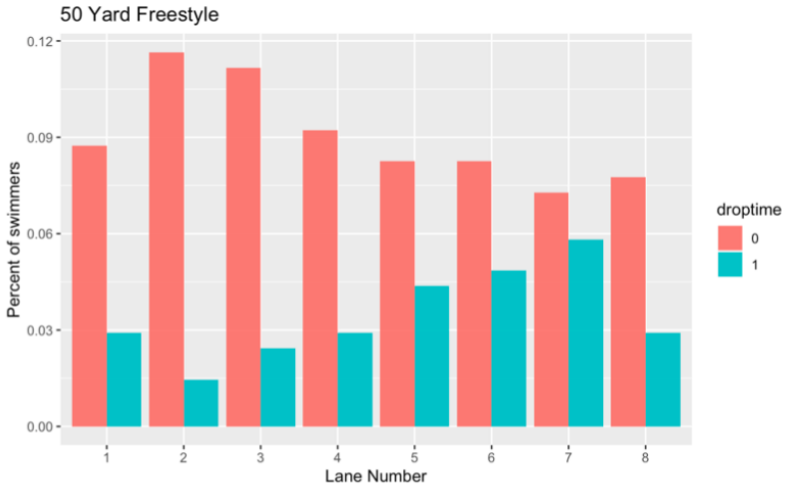

Figure 2.6: Distribution of the variable droptime (0 = No, 1= Yes) based on Lane number

In figure 2.7 representing the 1650 yard race, lanes 5 and 7 drop time more often than the other lanes regardless of gender. This raises the question of exploring more into the gender difference and will be examined next. Overall in the 1650 yard freestyle race, there is a much more even distribution of dropping time and not dropping time, as compared to the 50 yard freestyle event.

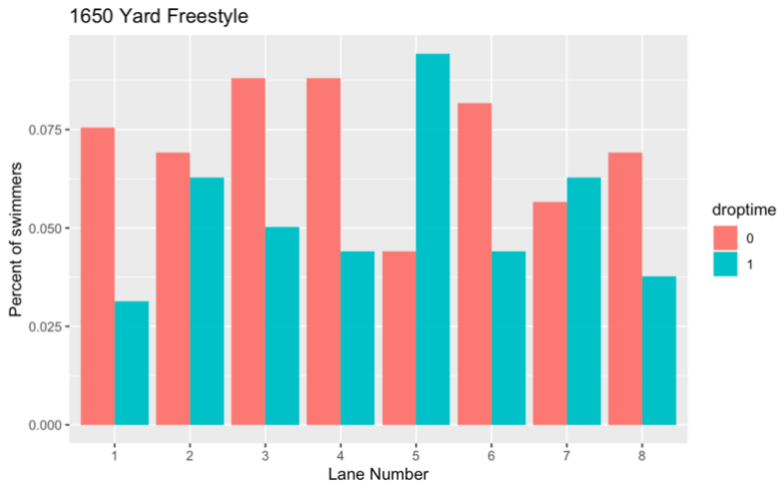

Figure 2.7: Distribution of the variable droptime (0 = No, 1= Yes) based on Lane number

In figure 2.8, all heats except for the first one (which is the slowest entry heat) have a higher percentage of people add time than those who drop time. Additionally, a pattern is found where as the heat number increases from 4 to 6, there are more people who drop time, but in the last two heats (and the fastest heats) there is a major drop off of people dropping time. The intuition that if a swimmer is entered faster and has a lower value for their entry_time, there is a smaller chance that they will decrease any more value from that time as compared to someone who has a higher value and more room for improvement. As the heat number increases, the eight swimmers in each of those heats have faster and lower value entry_time, so there is intuitively a smaller chance that they will decrease any more value from their time as compared to the swimmers in the earlier heats who have higher value entry_time.

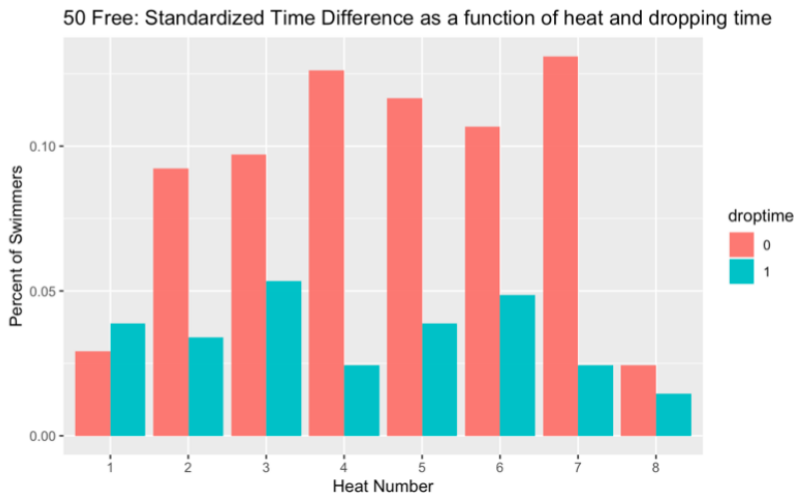

Figure 2.8: Distribution of the variable droptime (0 = No, 1= Yes) based on Heat number

In figure 2.9, looking at the 1650 freestyle race, a more even distribution of swimmers dropping time is noticed compared to those in the sprint 50 freestyle race. Only in heats 1 and 5, did a higher percentage of swimmers drop time than those who added time. We also notice that in the middle heats (3, 4, 5) a higher percentage of people drop time the later heat they swim in.

In both figures 2.8 and 2.9, a major discovery is that in the final heat, for both the 50 and 1650 race, the percentage of people who dropped time is the lowest in all the events. This formulates the idea that if swimmers are entered faster or higher ranked in the competition, as compared to the rest of their event, they are drastically underperforming in the race and have a smaller chance of dropping time.

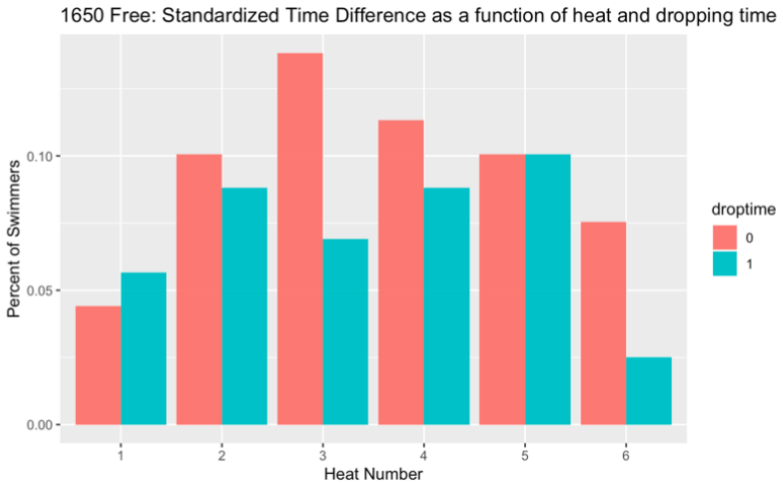

Figure 2.9: Distribution of the variable droptime (0 = No, 1= Yes) based on Heat number

In figure 2.10, the year 2016 has results where all four events had an average standardized time difference of a positive number meaning that most people added time. The next year, in 2017, three of the four events had lower average standardized times and additionally, the swimmers for those events improved their time except for in the women's 50 free. The men's races had similar trends for both events, whereas the women's events were not the same.

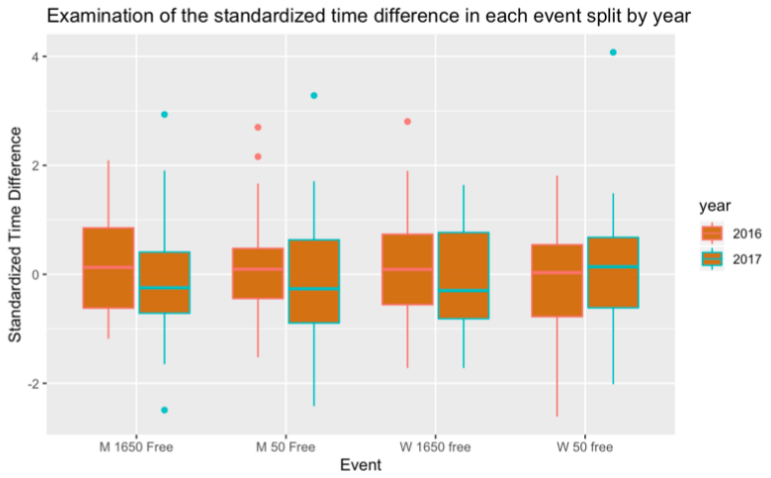

Figure 2.10: Boxplot distribution of Standardized Time Difference within each Event split by Competition Year

Figures 2.11 and 2.12 display the distinction between whether or not a swimmer makes finals based on their heat placement. Heat placement is a critical variable and being placed in a certain heat is a factor to consider when trying to predict the standardized time difference a swimmer has from their entry time.

In figure 2.11, there are no swimmers in the 50 yard freestyle from the first three heats that make finals regardless of if they dropped time or not. It was found that there are a greater amount of swimmers in the earlier heats that drop time as compared to the later heats. However, that being said, from 2.11 it is evident they do not drop enough time to be fast enough to be placed in one of the coveted top 16 ranks of finals.

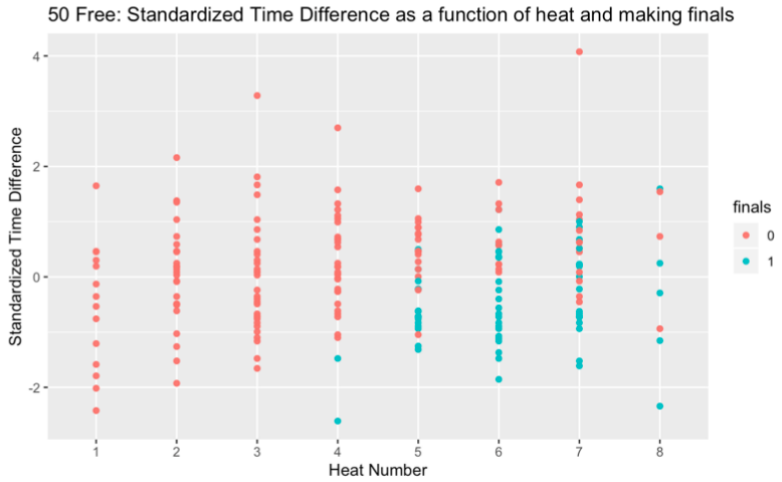

Figure 2.11: Distribution of the Standardized Time Difference within each Heat based on the variable Finals (0 = No, 1 = Yes)

In figure 2.12, unlike the 50 yard race, at least one person from each heat in the 1650 freestyle places in the top 16 ranks and makes finals. The spread between time difference is also more even throughout the heats, and there seems to be almost an equal number of people who do drop time and those who do not drop time. The large majority of those who do make finals are in those last two heats. This is expected because they were entered as the top 16 ranks and have a better chance of making finals, even though the probability of them dropping time is a different analysis.

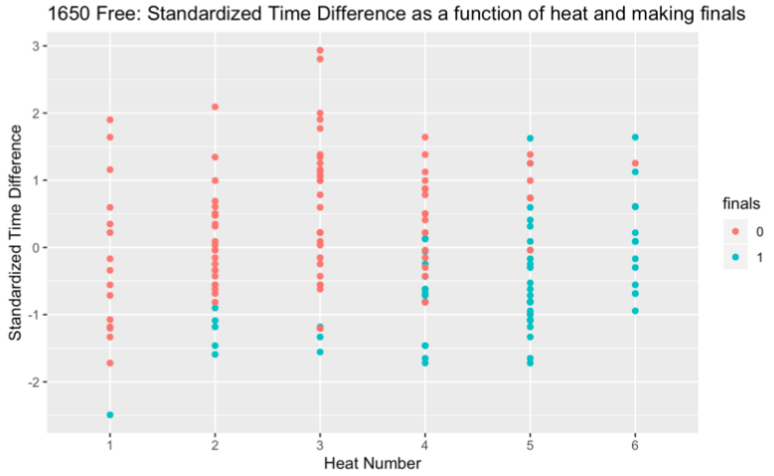

Figure 2.12: Distribution of the Standardized Time Difference within each Heat based on the variable Finals (0 = No, 1 = Yes)

In figure 2.13, the lane location (0=outside, 1=inside) of the 50 yard freestyle displays an interesting trend occurring in the inside lanes. In the outside lanes, there is a clear maximum density of swimmers producing results slightly to the right of the time difference = 0 value which means they are adding time and going slower. However, in the inside lanes, there are two peaks of density. One is slightly left of zero and the other is to the right of zero. This means that most people are either dropping a very little amount of time or adding around 0.25 seconds.

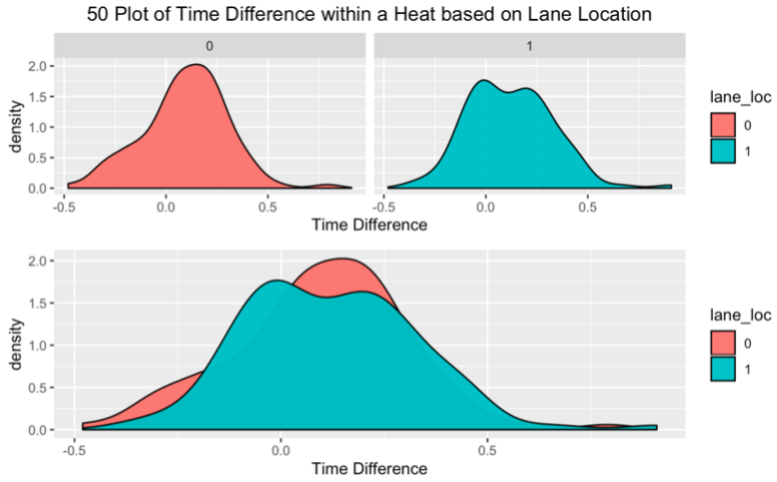

Figure 2.13: Density distribution of the Time Difference (time_diff) of all swimmers based on the variable lane_loc (0 = Outside, 1 = Inside)

In figure 2.14, the time difference result in the 1650 freestyle race is not greatly impacted by the lane location - inside versus outside lanes. Both of the lane locations have very similar spread and distribution centered around zero as the mean time difference.

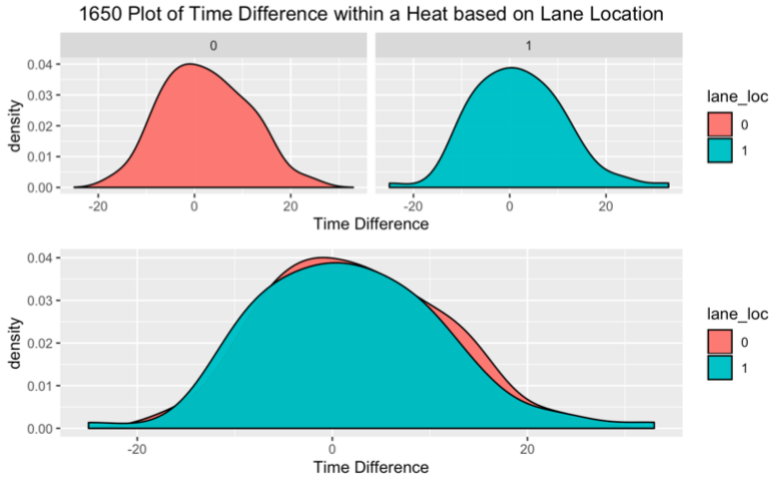

Figure 2.14: Density distribution of the Time Difference (time_diff) of all swimmers based on the variable lane_loc (0 = Outside, 1 = Inside)

In figure 2.15, the change in an average final heat place for each class between the two

years of interest is displayed. In most of the final heat places for each event from 2016 to 2017 there wasn't too much movement. A few things that stick out; the three biggest jumps were the seniors in the men's 1650 free race, the sophomores in the women's 1650 free race, and the freshmen in the men's 50 free race, are ranked much higher in 2017 as compared to 2016.

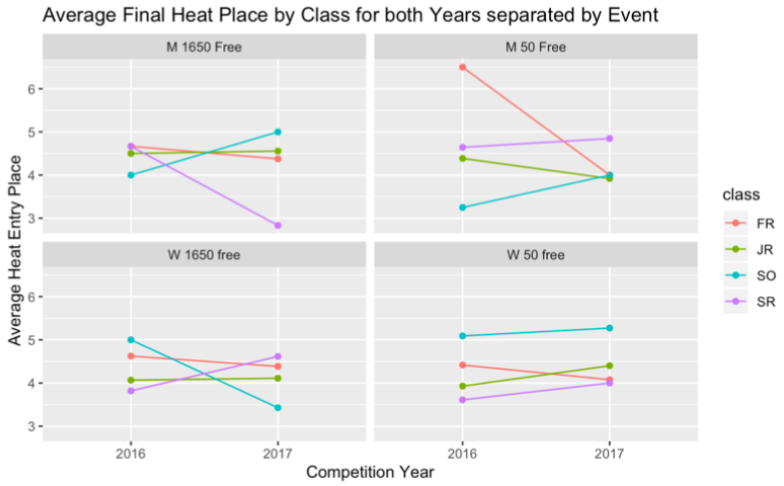

Figure 2.15: Comparison of Average Heat Entry Place for each Class and Competition Year

2.1 Top Three Schools

While it is an honor for any school and swimmer to receive an invite to the NCAA championships, the entire country watches in anticipation to see which school comes out victorious and wins the championship. As seen from the analysis above, there are many factors that impact how well swimmers do at the competition. I subset the top three schools based on the number of swimmers who were invited to the competition. It is interesting to compare how the swimmers in the top three ranked schools perform compared to the entire population in this study. The top three schools all have 20 or more swimmers invited. California and Texas both have 22 swimmers and NC State has 20 swimmers included.

The results coming out of the analysis of the top three schools proves that the idea of

depth a team possesses rather than having a few superstars will always win. The point system will favor teams that have a broader range of swimmers in many events that don't necessarily need to win each time, but need to have the most amount of people qualify for finals. In the finals round is where swimmers earn actual points to count towards the final score to determine a winner. The swimmers in finals are the best of the best and the driving factor for fans to jump up and down while screaming in excitement. Through the analysis, I found that there was only one swimmer who not only made finals both years in their respective event, but also won the event both years consecutively. While not an outlier in the analysis because of how close his entry time and final time were together, he is an outlier in looser terms because it is very rare to win two consecutive years.

Figure 2.16 displays the number of swimmers who added time (droptime = 0) and those who dropped time (droptime = 1) for the top 3 schools only. California has a greater percentage of swimmers who dropped time compared to the other two schools. Texas and NC State both have more people who added time as compared to those who dropped time within their school.

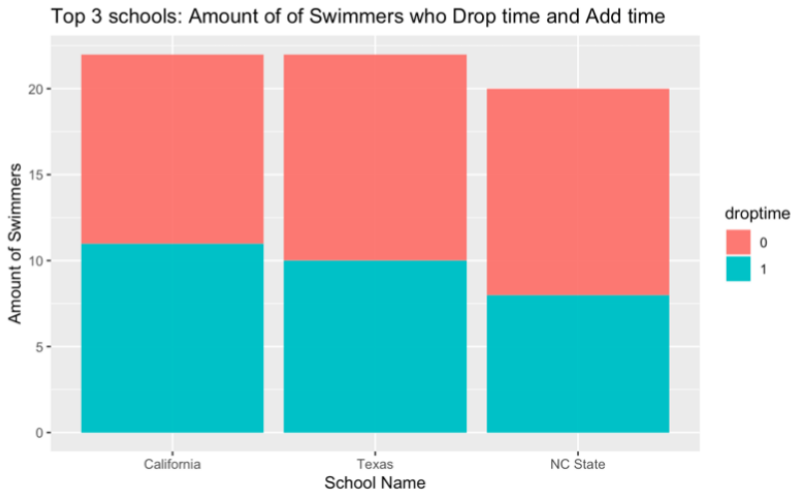

Figure 2.16: Count of swimmers from the Top 3 school split based on variable droptime (0 = No, 1 = Yes)

Figure 2.17 examines the results of the swimmers who attend one of the top three ranked schools in terms of the number of invited participants. Both women's races in 2017 performed better than in 2016 and the men had opposite results. The men's 50 free in 2017 proved that the younger the swimmer is, the better they performed, however, the 1650 race displayed reverse results with the more mature swimmers performing better in 2017. One explanation for these results is that in the short races, swimmers often feed off of adrenaline and the energy of their team. Given the swimmers in these short races had the most amount of teammates present at the meet, the energy, and adrenaline that surrounds them to sprint a twenty-second race is much greater than any of the non-top three teams present at the meet. The converse may not always follow true for the long fifteen-minute race regardless of the number of teammates a swimmer has at the meet because those races depend on much more than adrenaline.

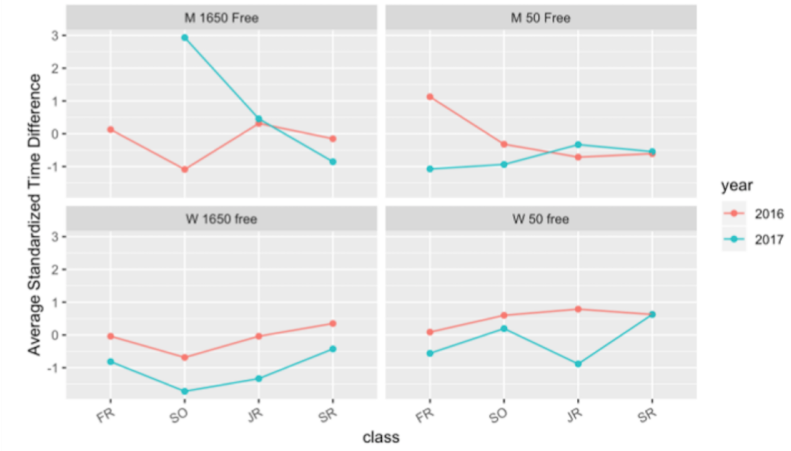

Figure 2.17: Comparison of the Average Standardized Time Difference for the Top Three Schools by Class and Event

Correlation analysis is an important part of any analysis. Typically, the variables must be all numeric; this study contains mostly categorical variables. A package called "lares" can be used on all types of data including categorical in addition to numeric [5]. For the numerical with numerical correlations, the Pearson correlation coefficient is being calculated as is standard, however, for the numerical with categorical, this package uses unique methods to solve for those. The corr_cross() function automatically converts categorical columns into numerical with *one hot encoding* (1s and 0s) and other smart groupings such as "others" labels for not very frequent values and new features out of date features [6]. This way more information and variables can be analyzed and discussed.

The results of Figure 2.18 follow, blue means a positive correlation between the variables on the left axis and the numeric percentage of that correlation on the x-axis. All of the presented correlations are 50% or higher and include the top 20 combinations of variables. Additionally, all of the correlations have a p-value smaller than 0.05, so they are all considered statistically significant as well. The variables entry_time and final_time have a 1-to-1 correlation and are present in most of the other variable combinations. Ultimately, regardless of the other factors and variables included in the study, the entry time and final time a swimmer displays are the most important factors to consider.

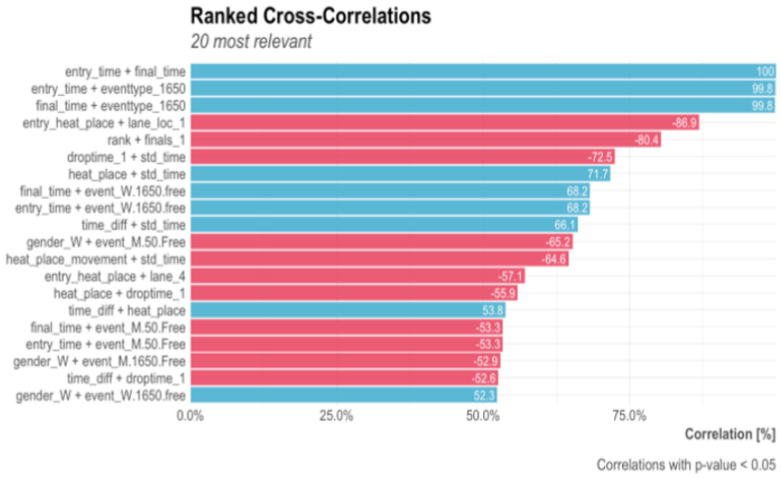

Figure 2.18: Top 20 Correlations

In Figure 2.19, the correlation analysis focuses on the top five variables correlated with the outcome variable std_time - standardized time difference. The variable droptime with level 1 associated with it, is the most important and correlated variable with std_time. It has a -72.5% correlation which means that as someone drops time the lower std_time they will have. This follows intuition because if someone is swimming faster, then they will most likely have a droptime level of 1. The second and seemingly more interesting correlation is heat_place and std_time; higher heat_place are associated with higher values in std_time. The other three variables that were important include time_diff which is the pre-modified version of std_time, so that is not worth looking into, heat_place_movement, and rank.

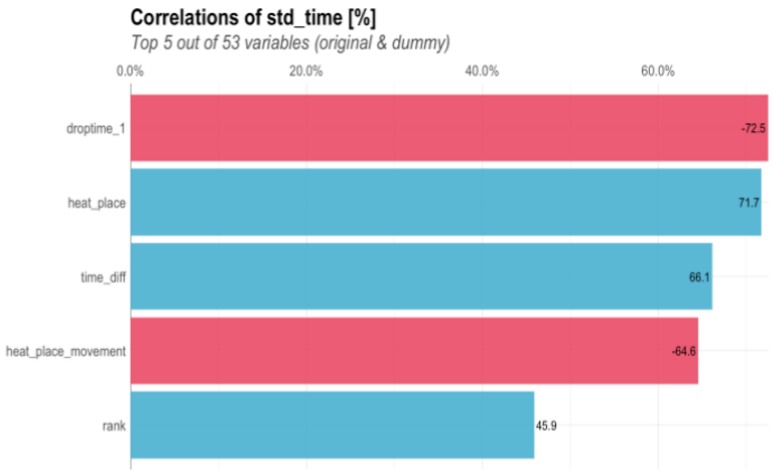

Figure 2.19: The top 5 correlated variables with the outcome varible std_time

CHAPTER 3

Variable Selection

In any model, variable selection is an important process to ensure the best fit such that accurate predictions can be made on the outcome variable. Choosing forward and backward stepwise selection are two standard methods to assist in choosing how many, and which specific variables are best to include in the predictive model. From the exploratory analysis, it is difficut to derministically state which variables are the most important to include without any reasoning besides intuition. Choosing to fit these model before regression will help provide concrete explanations and reasoning as to what factors are most significant in detemining the time difference a swimmer has between their entry time and final time.

3.1 Forward Stepwise Selection

In forward stepwise selection we begin with a null model meaning that we only include the intercept and no predictor variables. Once we only have the intercept, we begin introducing each variable while fitting p linear regressions. In this case, I will introduce a total of 14 predictor variables on my response variable of the standardized time difference ('std_time').

The forward stepwise process is essentially adding p linear regressions to the original intercept model to find the lowest RSS. Once those predictors are evaluated, we add onto the model with the subsequent two-variable model, onto the three-variable model, and so on.

The process explained above can be interpreted by the following algorithm:

Forward stepwise selection algorithm

1. Let M_0 denote the *null* model, which contains no predictors.

2. For $k = 0,...,p\text{-}1$:

· Consider all p - k models that augment the predictors in Mk with one additional predictor

· Choose the best among these p - k models, and call it M_{k+1}. Here best is defined as having the smallest RSS or highest R^2

3. Select a single best model from among M_0 ,..., M_p using cross-validated prediction error, Cp (AIC), BIC, or adjusted-R^2 [7].

To carry out forward stepwise selection I installed the R-package "leaps" and utilized the "regsubsets" function with "method = forward". In short, the regsubsets function is generically used to perform subset selection with methods for matrix and formula arguments. What is unique about the function is that it gives the user the ability to limit the number of p simple linear regressions it draws on with the argument "nvmax", which ultimately minimizes the computational workload [8]. The total amount of models would be

$$1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+2)/2$$

For the purposes of the data, I limited the p regressions to a "nvmax = 8" which ultimately gave me up to nine-variable models to analyze based off of the equation above (null plus the eight fitted models). In addition to saving computational power, the reason for limiting the maximum amount of variables allowed in the model helps avoid high model complexity which runs the risk of over-fitting with too many variables.

The output of received can be interpreted intuitively:

Each column represents the variable in the model. Factors are represented based on the "default" variable, for example, the default of 'class' is FR, or "freshmen". This means that

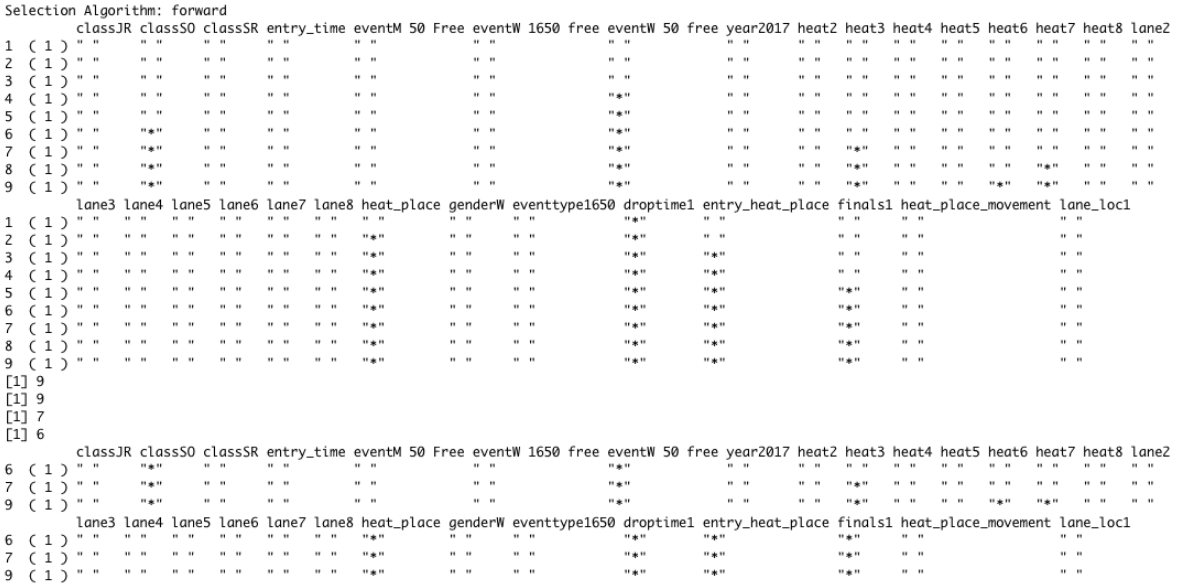

Figure 3.1: Significance of Variables in Forward Stepwise Selection

my model is using 'classFR' as a comparison basis for all of the other factors in the class variable. This principle holds true for all other factor variables such as 'event', 'heat', 'lane', 'gender', 'eventtype', 'droptime', 'finals', and 'lane_loc'. The rows represent the number of variables used in the model. So the one row will highlight the variable that has the most statistical significance out of all the other variables in predicting the standardized time difference. The way that the R output is "highlighting" the variables is by leaving asterisks under the variable column. This output becomes very intuitive and easy to digest when analyzing and interpreting the resulting p models.

After an initial look at the variables included in each of the models, the next step would be to interpret the plots and verification values to find the *best* model. As mentioned earlier, the algorithm uses RSS as a unit of measurement for the best fit model, however, other measurements can reinforce the final selection; Cp, BIC, and adjusted-R^2.

The next step in the analysis is to present the plots while highlighting the number of variables that optimize the verification statistics.

First, define what the various statistics are and how to interpret their significance:

RSS: Sum of Squares Residuals:

RSS is the Residual Sum of Squares. Residuals are the deviations predicted from actual empirical values of data. It is a measure of the discrepancy between the data and an estimation model. A lower RSS indicates a tighter fitting model of the data [9].

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

In Figure 3.2, the graph below shows the number of variables corresponding to each model (1 through 9), and our outcome on the y-axis is the RSS. A point was added to the graph that highlights which model results in the minimum RSS, indicating that it is the best model. It turns out that the best model, in this case, was the nine-variable model which yields the minimum RSS of 94.26246.

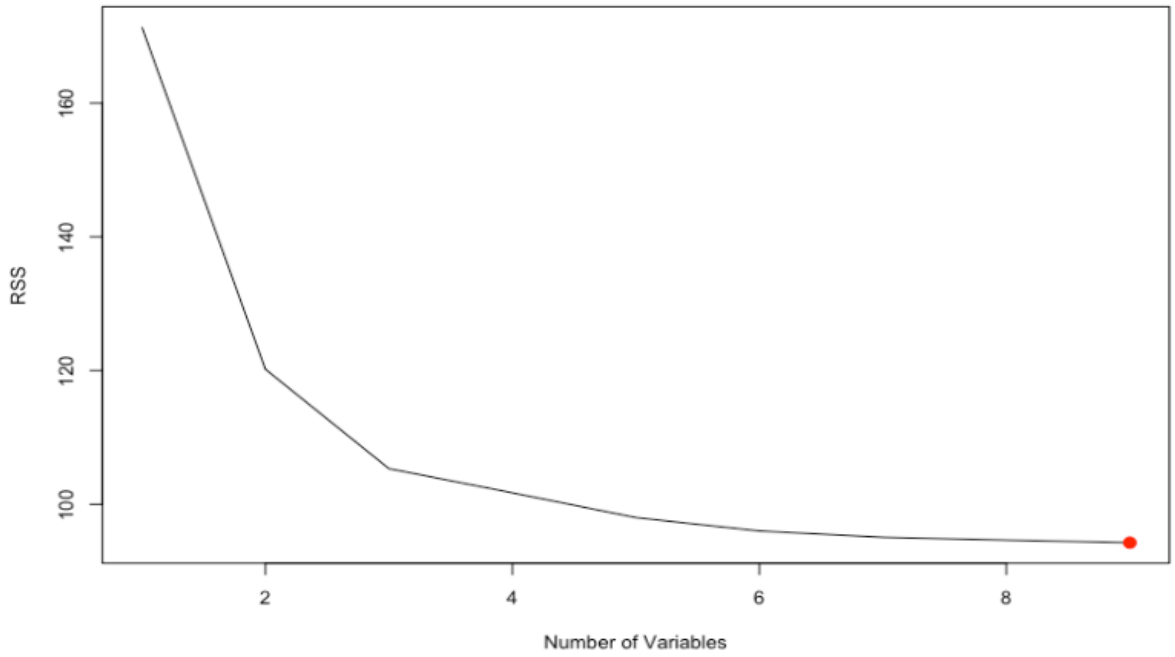

Figure 3.2: RSS value for each number of variables in Forward Stepwise Selection

C*p*:

The Cp statistic is similar to the RSS, however, it is adding a penalty to avoid overfitting and adjust for the fact that the training error tends to underestimate the testing error. This penalty is represented by the 2d$\hat{\sigma^2}$ and will increase as the number of predictors added to the model increases. In short, we want to select the model that outputs the lowest Cp, similar to how we did when looking at RSS, however, now we can adjust for overfitting [7].

$$Cp = \frac{1}{n}(RSS + 2d\hat{\sigma^2})$$

where $\hat{\sigma^2}$ is the estimated error variance.

In figure 3.3, the plot below is similar to the RSS plot from above, however, it is indicating that we can achieve the best model fit with only a seven-variable model. This follows the intuition and mathematical explanation of the Cp statistic because while adding on two more variables may decrease the RSS, the Cp statistic will add a penalty for two more predictors. At this point, the best model fit would be the seven-variable model with a Cp of -5.633192, instead of the nine-variable model because it will avoid bias and overfitting.

BIC: Bayesian Information Criterion:

BIC, or Bayesian Information Criterion, is similar to the Cp statistic because it is looking for the model with the lowest test error. The equation below demonstrates the similarity it holds with Cp, however, the penalty being used is $log(n)d\hat{\sigma^2}$ instead of $2d\hat{\sigma^2}$. Since BIC uses a log(n) penalty and log(n) $>$ 2, it is expected to select a model with fewer variables than Cp if n $>$ 7, which in this case, it is. Similar to the RSS and Cp statistics, looking for the model that will yield the least BIC value is needed.

$$BIC = \frac{1}{n\hat{\sigma^2}}(RSS + log(n)d\hat{\sigma^2})$$

In figure 3.4, the graph below shows a very similar downward slope as the previous two. The mathematics and harsher penalty due to the log(n) hold for my data because the model that yields the minimum BIC is the six-variable model with a value of -445.1179. As stated

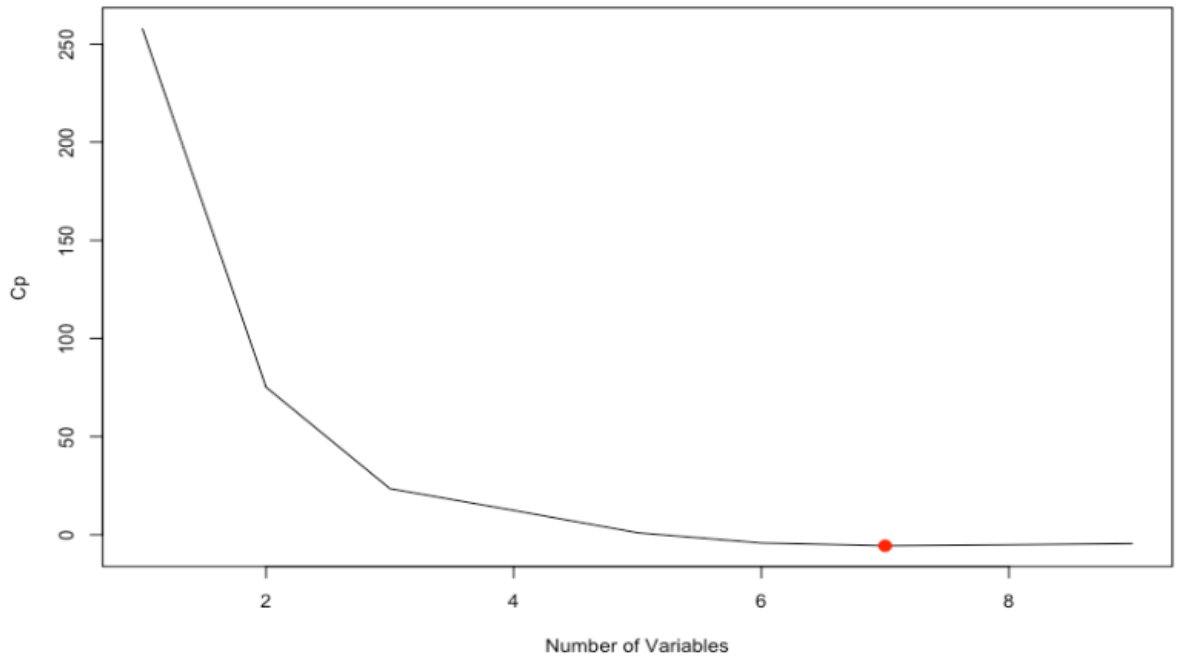

Figure 3.3: Cp value for each number of variables in Forward Stepwise Selection

before, this was expected because of the higher number of predictors included [7].

Adjusted-R^2:

The final modeling test statistic will be looking at the adjusted-R^2 value. The adjusted-R^2 is like the typical R^2, however, it is adjusting for the number of predictors added to the model. More specifically, it will only increase if the new terms improve the model more than the expectation. Because of this, the adjusted R^2.value will be lower than the original R^2 [10]. Unlike the previous three statistics, an examinatinfor the maximum value will occur. Since RSS is essentially calculated using the R^2 value, as R^2 increases, RSS decreases. Maximizing the adjusted-R^2 is equivalent to minimizing the RSS by adding more variables to the model.

$$AdjustedR^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

The plot below is essentially an inverse of the RSS plot. This is mathematically sound due to how each of the statistics are calculated. It is not surprising that the model with

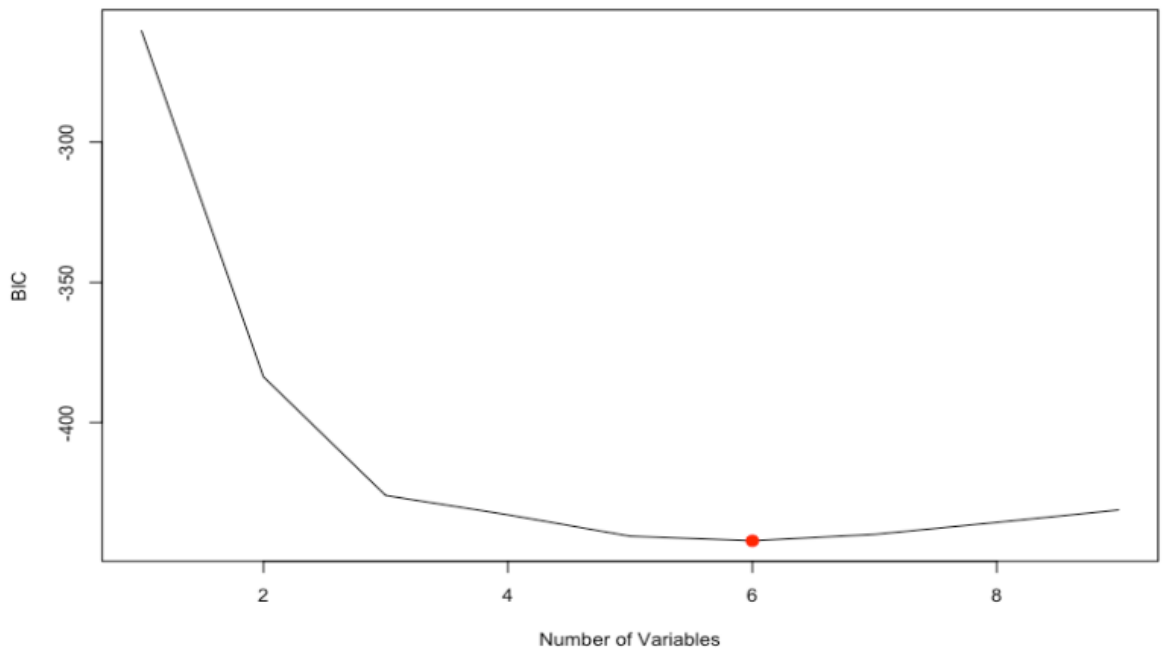

Figure 3.4: BIC value for each number of variables in Forward Stepwise Selection

the best, or maximum, adjusted-R^2 value is the nine-variable model with an adjusted-R^2 of 0.7322653.

3.2 Forward Stepwise Selection Results

After taking into consideration the best model fit statistics and verifications from above, after performing forward stepwise selection, the best model would be the six-variable model. While this model may not have the lowest Cp or RSS, it will have the lowest test error because of the low BIC, while only giving up 0.26 % of variance explained from the adjusted R^2 from the nine-variable model. These results will be used to be compared to the backward stepwise selection, and then followed by the mixed variable selection.

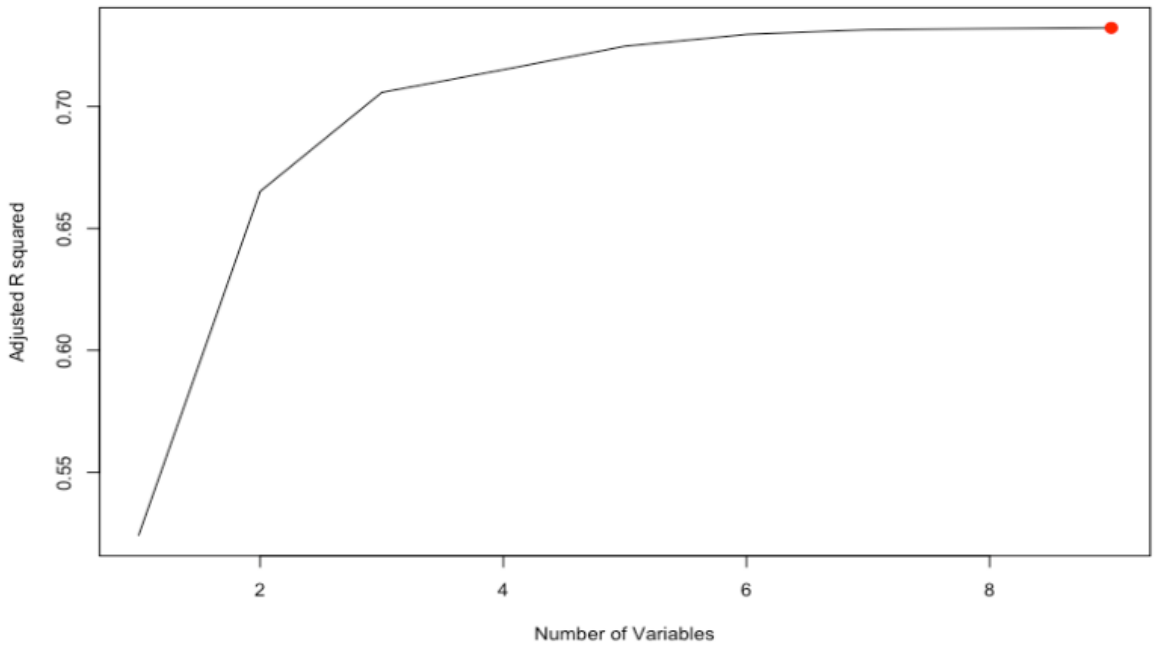

Figure 3.5: Adjusted-R^2 value for each number of variables in Forward Stepwise Selection

3.3 Backward Stepwise Selection

Backward stepwise selection holds true to its name in that it will have the opposite algorithm as forward selection. In backward stepwise, we start with the full model least squares model that includes all of the variables. We proceed by removing each variable that has the largest p-value and proves to be least statistically significant. These steps continue in a (p - 1) manner until the breaking rule is reached, one-at-a-time. A notable difference between backward and forward stepwise is that backward selection will not be able to be used if n number of samples is less than p number of predictors, however, the forward selection is a greedy method, therefore it can always be used under these conditions [7].

Backward stepwise selection algorithm

1. Let M_p denote the *full* model, which contains all p predictors.
2. For $k = p$,p-1...1:

Comparing Forward Stepwise Selection Model Statistics

	RSS	Adjusted R-squared	Cp	BIC
m1	171.36359	0.5240009	257.948730	-260.1582
m2	120.20600	0.6651797	75.172464	-383.6802
m3	105.34154	0.7057702	23.483462	-425.9600
m4	101.71604	0.7151074	12.388504	-432.8434
m5	98.01101	0.7247200	1.006310	-440.4869
m6	**96.00952**	**0.7295883**	**-4.222905**	**-442.1179**
m7	95.06534	0.7314976	-5.633192	-439.8253
m8	94.64236	0.7319414	-5.160952	-435.5530
m9	94.26246	0.7322653	-4.533111	-431.1212

Figure 3.6: Comparison of Forward Stepwise Selection based on the number of variables in each model

- Consider all k models that contain all but one of the predictors in M_k, for a total of *k-1* predictors.

- Choose the best among these k models, and call itM_{k-1}. Here *best* is defined as having the smallest RSS or highest R^2.

3. Select a single best model from among M_0,...., M_p using cross-validated prediction error, C_p, BIC, or adjusted-R^2

The same best-fit statistics will be used to evaluate backward stepwise, and the explanations of what they are, how they function, and their significance stays true to the explained section above.

3.4 Backward Stepwise Selection Results

After taking into consideration the best model fit statistics and verifications from above, after performing backward stepwise selection, the best model would be the nine-variable model. This model has the highest adjusted-R^2 value, the lowest C_p, and BIC values.

Comparing Backward Stepwise Selection Model Statistics

	RSS	Adjusted R-squared	Cp	BIC
m1	171.36359	0.5240009	257.94873	-260.1582
m2	120.20600	0.6651797	75.17246	-383.6802
m3	115.23335	0.6781414	59.21172	-393.2008
m4	111.03419	0.6890086	46.04478	-400.8500
m5	107.84789	0.6970915	36.53616	-405.5776
m6	105.29811	0.7034269	29.32658	-408.4108
m7	102.55267	0.7103504	21.41035	-412.1538
m8	100.44909	0.7154948	15.81240	-413.8188
m9	**98.53529**	**0.7201292**	**10.89996**	**-414.9401**

Figure 3.7: Comparison of Backward Stepwise Selection based on the number of variables in each model

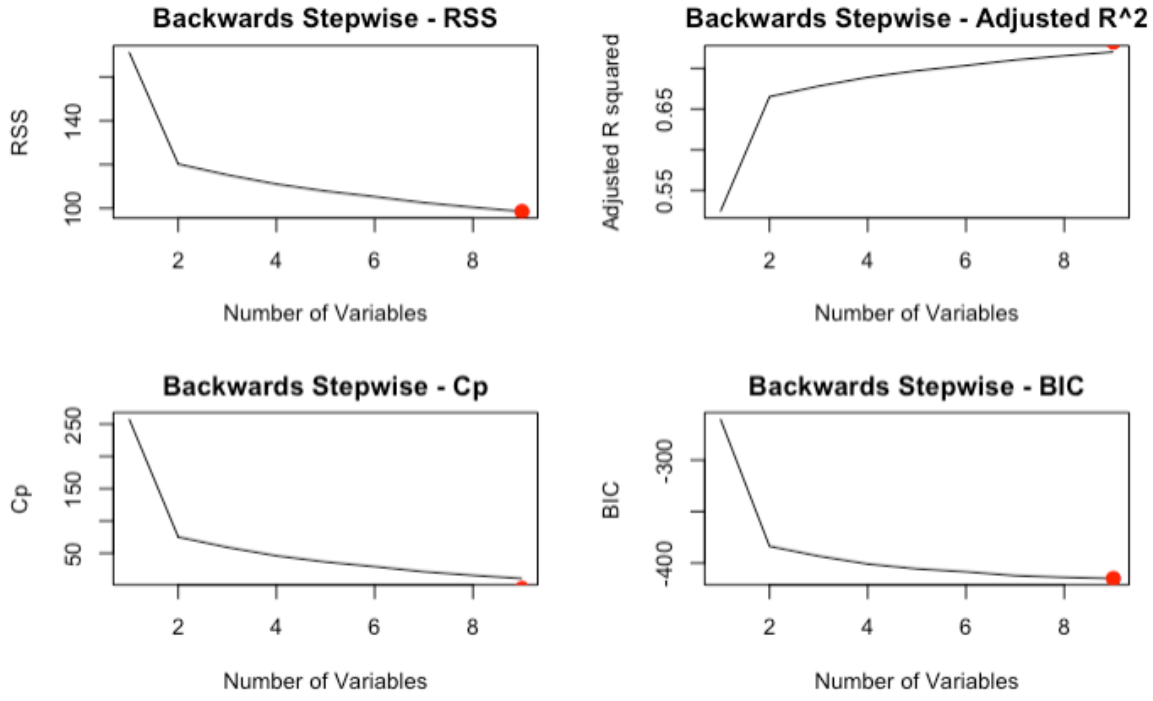

Figure 3.8: Evaluation Statistics of Backward Stepwise Selection

The plots above show that similar to the forward selection, a nine-variable model is best when looking at the RSS and adjusted-R^2. However, it also selected the nine-variable model as the best fit when looking at the BIC and C_p statistics as well. The table below shows that the difference between the nine-variable model for each statistic is greater than for the forward selections.

The coefficients of the statistically significant variables for the backward stepwise nine-variable model are as follows:

classJR	classSO	classSR	entry_time	eventM 50 Free	eventW 1650 free	eventW 50 free	year2017	
heat2	heat3	heat4	heat5	heat6	heat7	heat8	lane2	
lane3	lane4	lane5	lane6	lane7	lane8	heat_place	genderW	
eventtype1650	droptime1	entry_heat_place	finals1	heat_place_movement	lane_loc1			

(Intercept)	eventW 50 free	lane2	lane3	lane4	lane5	lane6	heat_place	genderW	entry_heat_place
-1.06296907	0.02488326	0.16258206	0.10296864	0.17421282	-0.04951498	0.05220147	0.36699786	-0.02235269	-0.12713550

Figure 3.9: Model Coefficients and Significance of Variables in Backward Stepwise Selection

3.5 Optimal Model Selection

Since forward and backward stepwise regression each contain a subset of p predictors used and result in different models, it is important to know which model is optimal. To do this, we have to look at a model related to the testing error rather than only being related to the training error, which both forward and backward are. To do this we have to partition the data to a training set and validation, or testing, set [7].

Once this was done, using the same regsubsets function utilized in forward and backward stepwise occured, however, this time it was not necessary to add the "method=", instead, fit the model on the training subset of data. The next step was to make a model matrix for the testing data. Finally, running a loop that extracts the coefficients from the best fit model and then forming predictions by multiplying the coefficients by the columns in the test model matrix. The result will be to compute the MSE's for each model and pick the model with the lowest MSE.

Figure 3.10 demonstrates that a three-variable model is an optimal model when looking at testing errors. The statistically significant variables that fall under this model are 'heat_place', 'genderW', and 'heat_place_movement'. The biggest discrepancy within these three different model selection techniques is that for both forward and backward selection 'heat_place_movement' was not statistically significant (even in a nine-variable model), and an interesting similarity is that 'genderW' is common among all three.

```
(Intercept)         heat_place             genderW heat_place_movement
-1.00983592         0.24095090       -0.04081589         -0.12478699
```

Optimal Model
MSE Results

	MSE
m1	0.5476345
m2	0.5477422
m3	**0.4445959**
m4	0.5580553
m5	0.4528691
m6	0.4583357
m7	0.4539675
m8	0.4664514
m9	0.4662606

Figure 3.10: Optimal Model Coefficients and MSE Results

CHAPTER 4

Modeling

4.1 Ridge Regression

In the least-squares, the goal is to estimate β_0,...,β_p while using the values that minimize

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2$$

.

Ridge Regression is similar to the least-squares, however, it is estimating values $\hat{\beta}^R$ to minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

For $\lambda \geq 0$ is a tuning parameter. The distinguishment between lambda being greater than 0 and $\lambda = 0$, is that when lambda = 0 then the tuning parameter goes away and we are left with the same equation and model as my least squares. It is important to be able to check the ridge regression model with the best lambda found by cross validation with the least-squares model ($\lambda = 0$) to see which model best fits the data and will have the best performance. The overall goal of both ridge regression and least squares is the same, to find the coefficient estimates that minimize RSS and best fit the data [5].

As λ grows and approaches infinity, so does the penalty, meaning that the model is less complex because the estimates start to converge to zero. When this happens, we can state

that there is high bias and low variance [5].

In ridge regression, it is also important to verify that the condition of the data matrix $\mathbf{X}$ being centered to a mean of 0 before performing ridge regression, is met. If this assumption is met then we will be able to estimate the intercept by

$$\hat{\beta}_0 = \bar{y} = \sum_{i=1}^{n} y_i/n$$

With all of this talk about the significance of the shrinkage parameter, λ, how is it possible to find the optimal value? The answer lies in cross-validation.

Selecting a Tuning Parameter for lambda: While there are many ways to perform cross-validation whether it be by k-fold or leave-on-out, I choose to do a 70/30 split manual split of the data to avoid the "black-box" of the integrated R functions.

To perform cross-validation and ridge regression I used the library 'glmnet' provided by R. The distinction between lasso and ridge regression within the function is that ridge will work on an α of 0, and lasso will work with an α value of 1. I then selected a grid of 100 lambda values. After performing cross-validation I found that the best lambda value yielded the lowest cross validation error of 0.09549433. When creating the model with my best lambda value I got an MSE of **0.2921816**. While this MSE was lower than the other arbitrary lambda values used, a limitation of this approach is that no variables are actually selected. Ridge regression does not perform variable selection because it does not lead any of the coefficients to 0, therefore it includes all of the features.

4.2 Lasso

Lasso is different from ridge because it not only uses a shrinkage parameter, but it also performs variable selection automatically because it will zero out the collinear variables.

The shrinkage parameter in lasso is

$$\lambda \sum \beta_j^2$$

[5].

The lasso coefficients minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid = RSS + \lambda \sum_{j=1}^{p} \mid \beta_j \mid$$

.

This harsher penalty is called the l1 penalty, which can force coefficients to zero when we have a high lambda value. In my case, I will not be choosing the lambda value arbitrarily, but by using cross-validation to give me a lambda value that will yield the lowest cross-validation error.

To perform cross-validation, I used the same function and methods as stated above in ridge regression, however, the only difference is that when I fit the model, the α level was set equal to 1 to indicate I was using lasso.

After using the same 70/30 split from above, I found the lambda value that yielded the lowest cross-validation error to 0.03055119. The MSE was **0.3463546**. This value went against my intuition that the lasso model would be a better fit due to the harsher penalty on coefficients.

4.3 Final Model

One of the disadvantages in ridge regression that occured when modeling the data compared to the lasso, is that it included all the predictors in the final model. Unlike lasso, the coefficients will not be zero, only start to converge to zero [7]. While this seemed like a disadvantage to the model in the beginning, a much lower MSE is achievable when performing the ridge model when compared to lasso. This might be because the lasso penalty was

too harsh. The final ridge regression model can be expressed as the following, with the accompanying coefficient outputs.

(Intercept)	classJR	classSO	classSR	entry_time	eventM 50 Free
-0.18425	-0.00377	0.14125	-0.00595	0.00005	0.02685
eventW 1650 free	eventW 50 free	year2017	heat2	heat3	heat4
0.00462	-0.09996	-0.02543	-0.00008	0.13570	-0.02106
heat5	heat6	heat7	heat8	lane2	lane3
0.00074	0.12981	0.17524	0.19321	0.02569	0.02889
lane4	lane5	lane6	lane7	lane8	heat_place
0.01468	-0.07109	0.01182	-0.04584	-0.05195	0.12242
genderW	eventtype1650	droptime1	entry_heat_place	finals1	heat_place_movement
-0.08010	0.06204	-0.82110	0.00069	-0.27280	-0.09844
lane_loc1					
-0.00699					

Figure 4.1: Final Model Intercept and Coefficients

Discussion

The model outputs provide more concrete answers that tie the results of the exploratory analysis together. The results are consistent with what the analysis above found. The optimal variable selection model showed that heat_place and heat_place_movement are significant and this was displayed as well in the exploratory analysis. It was interesting that the optimal model concluded that a three variable model was sufficient to predict the outcome even though the data contains 20 variables total. The class a swimmer is in was not as deterministic as initially believed. Additionally, the lack of signifcance lane number had on the outcome time difference was a new insight, that intuitively was believed at first.

CHAPTER 5

Conclusion

The objective of this research paper is to discover insights and find the best predictive model that is able to identify the most important variables that impact the time difference between entry and final time at the competition. I first began with exploratory analysis and extracting valuable insights from each variable to understand the differences and relationships between each other. The question to be answered by the exploratory analysis is why do most people swim slower at the nation's fastest collegiate competition and what factors impact this result the most. There were many takeaway ideas discovered from the data analysis as the exploratory portion was the largest chunk of the project.

First, according to the plots, on average, the results in 2017 were better than 2016. This follows prior intuition that regardless of how many people add time and go slower, each year the competition gets slightly faster. The positive and negative time differences for each of the 8 events highlighted that the shortest 50 yard race has the greatest percentage of people adding time. In addition, the women's events had the top three percentages of time added. In contrast, the longest 1650 yard race had the most amount of time dropped. In this case, it was the men who dropped more time than the women. All other factors equal, the class a swimmer is in does impact how well they do at the NCAA competition. While all classes on average add time, the women's 1650 freestyle shows a clear negative linear relationship. This result holds true for the men's 1650 freestyle as well, not including the freshmen. The more mature a swimmer is and the more exposed they are to these types of high-level and nerve-wracking competitions, has a direct effect on how well they execute their performance.

This idea holds true especially for the longest event of the 1650 freestyle.

Next, the results of heat and lane number placement impacting time difference proved insignificant. Singularly, lane number did not have an impact on whether a swimmer added or dropped time. While inside lanes place higher within their heat, thus having a better chance at making finals, they do not perform as well when looking at their time difference between entry time and final time as compared to outside lane. Outside lanes perform better in terms of time difference, but do not move up the ranks enough to have a better chance of making it into the finals round. All other factors constant, heat number did not have a large impact on the time difference. The last heat always had more people swimming slower than their entry time, and the first heat had more people swimming faster than their entry time. This means that if a person is entered faster, they have a smaller chance of going any faster compared to those people entered slower. This follows my intuition because there is an unmeasurable pressure element that falls upon the faster swimmers to make the finals round. Oftentimes, the swimmers in the first few heats do not feel the same pressure and are lucky to have just been invited to the competition.

Finally, the results surrounding the heat and lane placement found an overall insignificant impact on time difference with the exception of the tail ends of each variable. These results regarding the first few and last few entry heats play with the idea of regression to the mean. **Regression to the mean** is the idea that if a small group on either side of a normal curve is sampled, the mean will be different than the population mean and with regression it increases/moves the small sample mean of that group closer to the population mean [11]. This idea is prevalent in the data in multiple ways at various levels because the swimmers who competed at the NCAA competition in the first place are the top swimmers in the nation and fill in the far right side of the curve. Deeper application is within all the swimmers invited to the meet, there exists a smaller subset of swimmers who make finals. Only 16 swimmers in each event qualify for finals, and being fast enough subsets them to the far right of the curve again. Although not explicitly examined, regression to the mean is an important idea that

occurs within data due to the nature of the size of invited swimmers to this competition. In future work, this idea can be more heavily explored with analysis and an exploration into the group of people who are the fastest and make finals compared to those who do not.

Regardless of all the variables and models tested in this project, we must realize that swimming and sports in general have incalculable factors that go into play with how an athlete will perform. Pressure, hype, and energy. The butterflies in your stomach as you step up onto the cold block, the pace you hit in warm-up and the energy of the cheering crowd will always be something that impacts how you perform that cannot be measured in a dataset. This paper attempts to provide reasoning and results as to why certain outcomes happen, but we must not forget about the measureless data. So even with all of the numeric and categorical data available, there will always be alternative reasons for why a swimmer did or did not perform well and no amount of data will tell the full story.

5.1 Future Analysis

From our findings, it is evident that most people, regardless of how the NCAA results are divided, add time. Now that this idea has been executed and discovered, this information is beneficial to athletes, coaches, and teams who are searching to find ways to prepare before coming so that they have a higher chance of success at the competition.

A starting recommendation is that more numerical data could be added to the data sets to allow for more precise analysis and predictions. This includes split times of the races at recurring intervals such as every lap (25 yards) or halfway points. These additional split times could be used to examine if people swim faster in the first half or second half of the race and how that impacts the final time. In addition, having 25 yard splits can be useful to recommend more accurate race strategies before even beginning the race. Further, having information on the date of qualification and time progression results for each swimmer would be exceptional variables to examine. Swimming is one of the longest season sports that

covers two academic quarters as compared to other sports that land in only one quarter. Therefore, utilizing the results each swimmer posts throughout the season will allow for a deeper understanding of when they perform their best; are people early-season, mid-season, or postseason swimmers. Another important measure that could be taken, is to use future data (the following years, 2018 - 2019) and see how well these current models perform and predict the future results. This would greatly strengthen the confidence in these models and allow future results to possibly have more positive results. Finally, rather than performing the analysis on just two events, expanding the project to cover other events including different strokes and distances would be beneficial. This project looked at just freestyle events, however, I am very curious to see how other stroke results would look like.

Incorporating these aspects into the data would allow for further research and analysis to be done and the utilization of the models would produce higher accuracy in recommendation for how coaches should train swimmers and when the optimal performance period occurs.

REFERENCES

[1] Patrick O'Rourke. Odds of a high school swimmer competing in college 2020. https://scholarshipstats.com/swimming, 2020.

[2] Jared Anderson. Ncaa refresher: How to qualify for the ncaa division 1 championships. https://swimswam.com/ncaa-refresher-qualify-ncaa-division-championships/, March 2014.

[3] Stephanie Glen and C. H. Goulden. T test (student's t-test): Definition and examples. https://www.statisticshowto.com/probability-and-statistics/t-test/.

[4] Paired sample t-test. https://www.statisticssolutions.com/manova-analysis-paired-sample-t-test/.

[5] Bernardo Lares. R correlations. https://github.com/laresbernardo/lares, 2020.

[6] Find insights with ranked cross-correlations. https://datascienceplus.com/find-insights-with-ranked-cross-correlations/, August 2019.

[7] Witten Daniela Hastie Trevor Tibshirani Robert James, Gareth. An introduction to statistical learning with applications in r. pages 79,207–219. Springer, 2013.

[8] Thomas Lumley. regsubsets. https://www.rdocumentation.org/packages/leaps/versions/2.1-1/topics/regsubsets.

[9] Residual sum of squares. https://en.wikipedia.org/wiki/Residual _sum_of_squares, June 2020.

[10] Multiple regression analysis: Use adjusted r-squared and predicted r-squared to include the correct number of variables. https://blog.minitab.com/blog, June 2013.

[11] William M.K Trochim. Regression to the mean. https://conjointly.com/kb/regression-to-the-mean/.