**Introduction**

The purpose of this project is to use survival analysis to study the career lengths of right handed, left, and switch players. Survival analysis is a method for analyzing data where the outcome variable is the time until a specific event of interest (Despa). The events can be death, marriage, divorce, decay rate, and career length. During survival analysis, subjects are followed over a certain amount of time and the main area of interest is the time at which the event occurs. An important thing to note while conducting survival analysis, is that data or observations can become censored. Censored refers to when data regarding survival time is incomplete (Despa). Right, left, and interval are the three types of censored data that exist within the realm of survival analysis. When conducting survival analysis, the survival and hazard function are an essential component. Simon Despa describes the survival function as giving the probability of surviving up to a certain time. Additionally, the hazard function gives the potential that an event will occur, per as specific time unit, given an individual have survived up to a certain point (Despa).

**Question**

The question we are trying to answer is who has the longest career in baseball right handed, left handed, or switch players. Being able to predict which players have the longest career, can help owners and coaches select the players, who will provide the greatest benefit to the organization over the long run. After the work was complete on this original question we shifted to explore how accurately a player’s handedness actually predicts their lifespan in the major leagues.

**Data**

To construct our model, we took data from the following files Masters.csv and Batters.csv. The data we took from Masters.csv file included playerID, name, weight, height, bats, throws, debut, and final game. From Batters.csv, we took selected the following data playerID, last year, and total seasons. Merging the scraped data together we created the table “final”, which we used to develop our models.

After constructing our data file, our first step was to remove any bad data points. Mainly, we removed data that was incomplete. Additionally, we added a boolean to determine whether a player is alive. The boolean mainly looks for players for last year in the MLB was 2016. If the last year is 2016 the player is assigned a 0, meaning they are still alive or their career has not yet ended.

Now that we a final data set we can begin the survival modeling process. Essential to survival analysis modeling is the creation of the Kaplan Meier chart and the survival curves. The Kaplan Meier chart takes the number of players in our data set and estimates how many players will remain after a certain time period. In these data sets the measure of time is the number of seasons of baseball that a player plays. All predictions is based on the totalSeasons variable that was mutated into our dataset, and is a count of the amount of seasons that a player plays in the MLB.

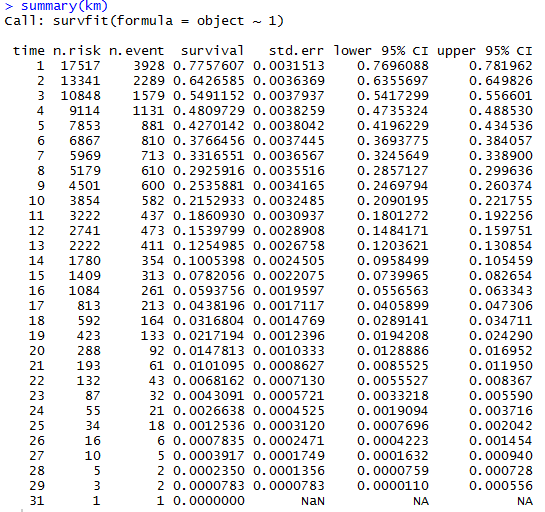


Figure 1: Kaplan Meier chart generated using model

As you can see in figure 1 it takes the total amount of observations in our data and predicts how many players will remain after n seasons. In addition, it allows predicts the percentage of players that will survive after *n* number of seasons in MLB. In Figure 1 above it can be seen that about 22.5% of all first year MLB players will not survive their first season. Another interesting point, one player from our dataset, Deacon McGuire, lasted a whopping 31 seasons of baseball.

Another important step in survival analysis is the creation of the survival curve. The survival curve displays the survival probability of an observation or a specific time measurement, in our case it is the number of total number of seasons a player will play. Using our model, we were able to generate three survival curves. The first survival curve examines batters swinging hand and the examines whether a player throws left or right. Finally, the last model is a combination of swinging and throwing together.

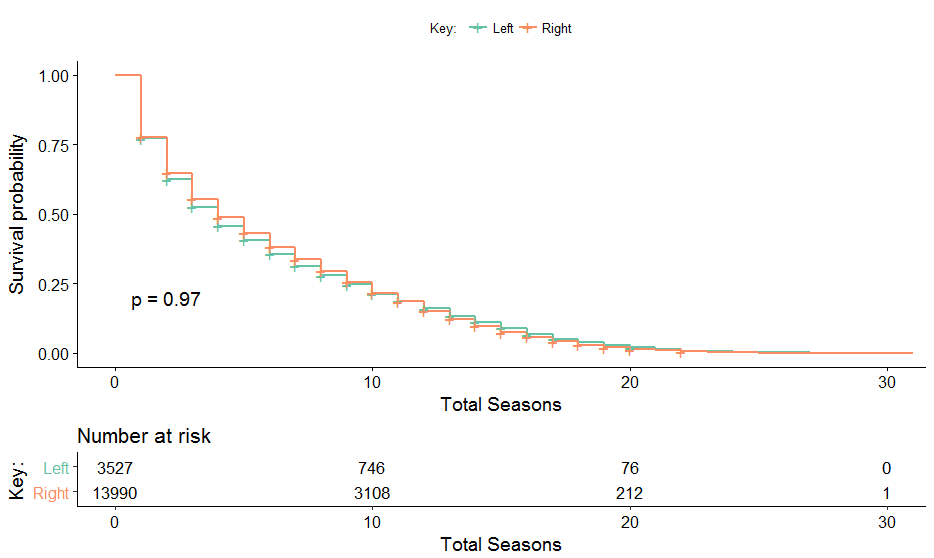


Figure 2: Survival curve for players who throw left or right handed.

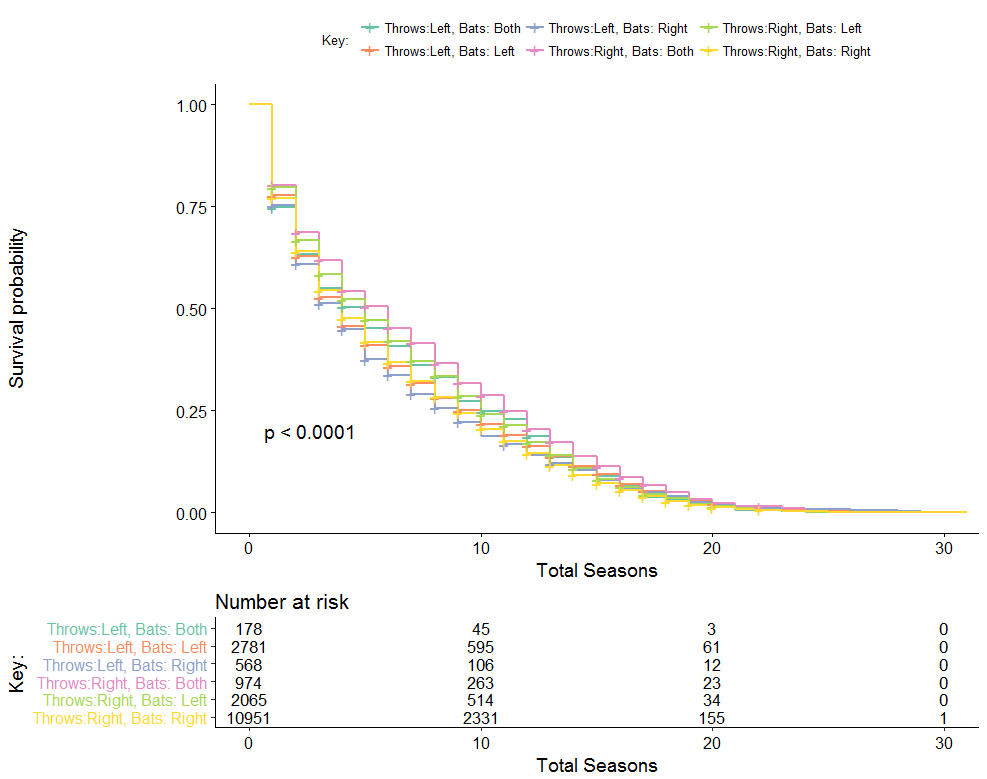
Figure 2 displays the survival curve the measures the probability of a left hand or right-handed player surviving. As, you can see from the graph right handed throwers have a higher chance of survival or left-handed throwers for the first ~10 years of their career, but after that you are more likes to survive ever-so-slightly as a lefty. However, this data is insignificant because of the extremely high p-value. The data does not fit this model very accurately, therefore it cannot be conclusive that a right or left-handed thrower will survive longer.

Figure 3: The survival curve for players with respect to batting and throwing handedness

The data in Figure 3 is far more conclusive to who will survive longer than that of Figure 2. With an extremely low p-value, it can be concluded that this model fits the data points extremely well and is a much more accurate predictor. The conclusions that can be made from this is that right-throwing switch hitters have the most consistent probability of surviving longer than other players.

**Model assumptions**

Our model treats players, whose last year was 2016, as still currently playing. Due to the dataset that we were working with, the BaseballReference.com archives, data points only extend to the end of the 2016 season, so these were the players that we considered as “alive”, or still playing today. This means that our data is missing potential points that it could have gotten from the past couple years of baseball, and is only able to predict the career length of the rookie players of 2016. This indicator, or censor, was mutated into the dataset under the isIn column, an is a zero if the player is still alive, and a one if they are “dead” or stopped playing in the major leagues prior to 2016.

**Model Limitations**

The fundamental limitation that we encountered is that every log in the Batting.csv dataset counts as a full season. This means that if a player steps on the field for a single game for a team and strikes out once, our model will treat their data the same as one of Hank Aaron’s peak seasons at the height of his career. As with many professional athletics, it is very common for athletes to get injured and take portions of a season off. Our model does nothing to handle these cases due to the data we are given.

**Model Improvements**

Our model could have been improved if we considered thee more aspects. The first thing we could not consider is continuously changing statistics. All of the data we included in our model was constant data. Additionally, we did not take into account switch throwers. If we took switch throwers into account it possibly could have greatly affected the survival curve for throwing, this is because there were very few, all of which extremely inconsistent, data points for switch throwers. Lastly, our model could have been improved if we considered players who had incomplete data files.

**Prediction and analysis**

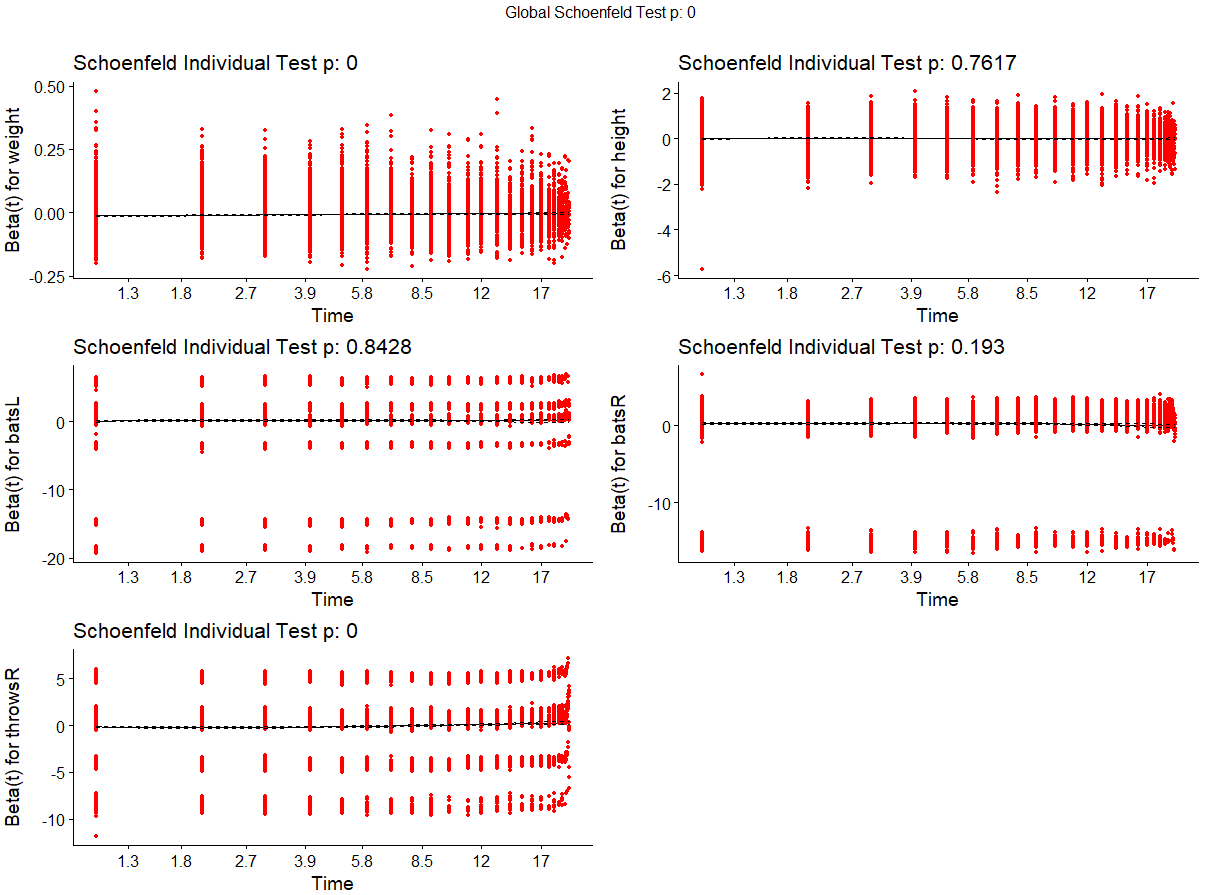
We took our data and fit it to a Cox model where we tweaked input to find a best fit for the different data we were attempting to consider. We built our initial model off of height, weight, batting, and throwing handedness. When removing height and weight we found that it decreased our overall p-values, despite those being the data points with the highest p-values, therefore we stuck with our initial model including all four. 

Figure 5: A plot of the hazards from our model

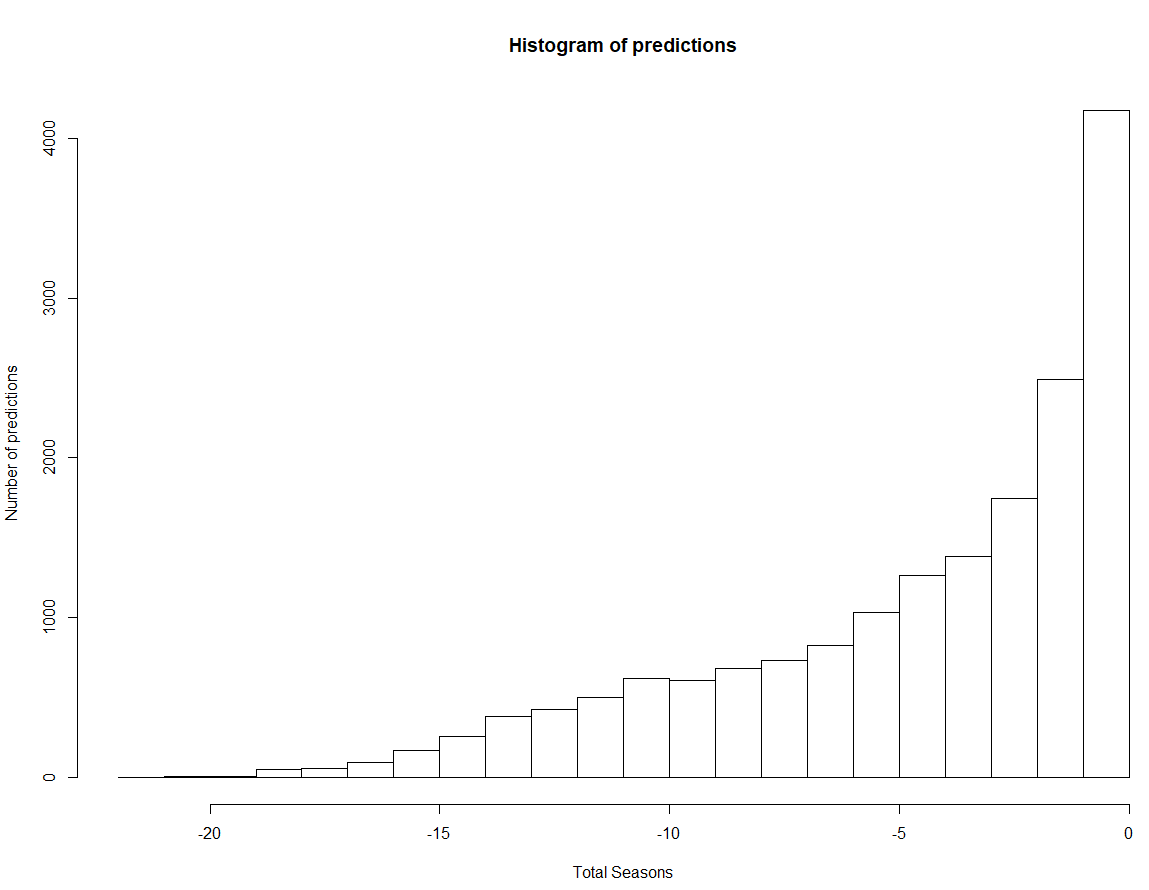
It can be seen in Figure 5, especially in the bottom three graphs, that there are systematic departures from the horizontal axis. This suggests non-proportionality in our hazards and can lead us to reject the null hypothesis that our data is proportional. 

Figure 6: Predictions versus total seasons incorrectly predicted

The histogram in Figure 6 above shows that our model continuously predicts the data short, meaning that it expects players to die before the actually do.

**Conclusion**

Using our model, we were able to generate three different survival curves. The first survival curve mainly looks at different batters. The second survival curve compares left and right-handed throwers. When we look mainly at batters, switch hitters have the highest chance of survival. Similarly, we compare left handed and right-handed throwers to one another, we see left handed throwers have greater chance of survival compared to right handed throwers. Finally, the last survival curve compares different combinations of throwers and hitters against one another when taking into account a player’s throwing and batting hand, we see that players who, throw right and can bat using either hand have the greatest survival rate.

Works Citied

*Despa, Simon. StattNews #78: What is Survival Analysis? New York: Cornell University: Cornell Statistical Consulting Unit. https://www.cscu.cornell.edu/news/statnews/stnews78.pdf*