**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, 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.

**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 file final.csv, which we used to develop our model.

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.

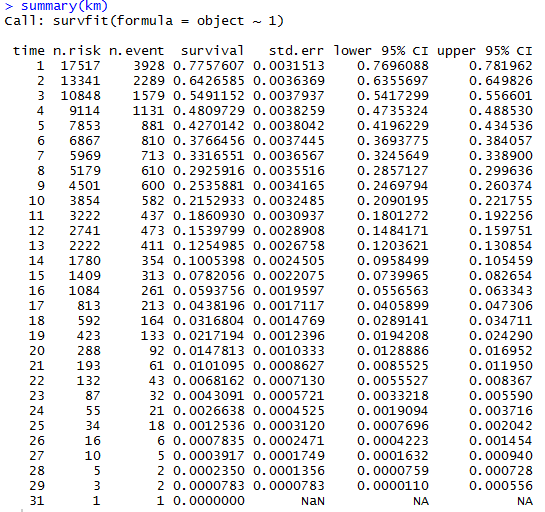


Figure 1: Kaplan Meier chart generated using model

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* amount of seasons in MLB

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 curves looks at batters swinging hand and the second looks at whether a player throws left or right. Finally, the last model is a combination of swinging and throwing together.

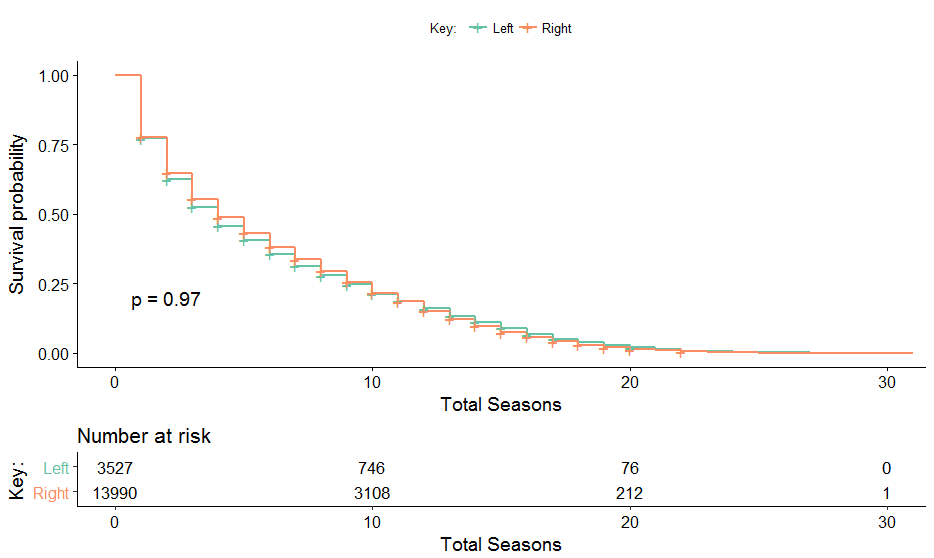


Figure 2: Survival curve for players who throw left or right handed.

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.

**Model assumptions**

Our model treats players, whose last year was 2016, as still currently playing.

**Model Limitations**

One limitation of our model is that it does not take into account whether or not a player plays a whole season. If a player only played one game our model, counts that one game as a whole season.

**Model Improvements**

Our model could have been improved if we considered thee more aspects. The first thing we failed to consider is continuously changing statistics. Most of the data we included in our model was constant data, and there was not much variation among it. 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. Lastly, our model could have been improved if we considered players who had incomplete data files.

**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 players 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*

left or right hand have the greatest survival rate.