Assignment 3-Markov Chains in MLB

George Brown

Data Preparation

For these models we will look at the 2017 event data to analyze ten thousand simulations in the Major League Baseball season. We will use Markov Chain transition probabilities to estimate the end of inning runs scored associated with each of the 24 states based off of the simulations for all MLB teams as a hole. We will then do the same thing except we will do it just for the Houston Astros and Los Angeles Dodgers 2017 season. We will only look at the Astros and Dodgers and compare the two teams expected runs by each state over the 2017 season.

The last model we look at is we have to make 16 Markov models for individual players in each lineup. We also are supposed to do it in a particular order. The order for the Dodgers is Chris Taylor, Corey Seager, Justin Turner, Cody Bellinger, Yasiel Puig, Joc Pederson, Logan Forsythe, and Austin Barnes. For the Astros the starting lineup is Jose Altuve, Alex Bregman, Carlos Correa, Marwin Gonzalez, Yulleski Gurriel, Brian McCann, Josh Reddick and George Springer. We will look at each player in the starting lineup and then pick the probable winner of game 7 between the Astros and the Dodgers. At first glance of the data it is very difficult to tell anything, so we will have to make a model of the expected runs per state to better analyze the data. There is not much data to understand until we model and then we can compare the Dodgers, Astros, and the MLB as whole.

Review Measurement and Choices for Markov Models

Our goal in creating these Markov chains is to predict based on event data what will happen and how many runs each team will score at each particular state. There are a couple issues with these Markov chains. One is while we are creating these we are just doing them over one year. While this is a lot it could be better if we did them over a couple years. Another issue is the DH. In the American League there is a DH and in the National League there is not. This makes the states slightly different because the ninth hitter in the AL is going to be a much better hitter than the pitcher is in the NL.

A couple choices we make are we choose to run 10,000 simulations across the MLB to give us a big sample size for the simulations. We also choose to look at just the 2017 event plays. Finally for this first model we want to look at the expected runs scored at the different states. There are a total of 25 different states including the 25th one which is three outs. The other 24 depend on the number of outs and the runners on base. For instance one state would be with nobody on base and nobody out, while another state would be a runner on first and no outs. The second model will be just for the Astros and Dodgers, so we will keep the methods the same. We then choose to make induvial models on each player in the game. The one big decision we make about simulating the game is that the pitchers will always strikeout. While this is not perfect more times than not a pitcher does get out. On average they get out more than any other position player. Another important decision we make is we do not factor in home and away. Being the home team in sports can be an advantage. I do not think being at home matters as much in the MLB as it does for college basketball. For game 7 the Dodgers were the home team and we could have given them an advantage in our model because of that. However, we decided just to base the simulations off of the events in the 2017 year.

Implementation and R Programing

There were many steps we had to take when implementing these models for the entire MLB. The first was we had to convert the play by play into CSV files to be able to import them into R. After we imported each individual team we combined all 30 teams into 1 play by play set defined in R as “MLB”. Then we defined all the variables in the play by play set. We then create indicators for runners at each base. The next step was to create a variable to indicate the state of each play and the next state. After we converted the matrix to a probability matrix we then made it into a Markov Chain. We then calculated the expected runs at each of the states. The first number of the states was the number of outs and the next three represented if runners were on first, second, and third. Finally, we calculated how many runs you would expect to score at the end of an inning depending on what the current state was and made this into an expected runs table.

For the Astros and Dodgers we did the same method as we did for the entire MLB except we used a subset of the Astros and Dodgers. We removed the rest of the teams and only focused on the at bats of the Astros and Dodgers. We wanted to focus only on the two teams who played in the World Series. We also found an expected runs by each state for each of these teams.

For the third step the modeling began to get more difficult. We created individual models for each player on the team. The first step wat to put them into a matrix and then we converted them into induvial Markov chains. We then put them in order that they would be batting. We created a pitching model Markov chain assuming they would always strikeout. The final part of the modeling was we had to simulate game 7 of the World Series and predict the probability the Los Angeles Dodgers would win the game. We had to run 10,000 simulations based off of the batting order for that game using the individual player’s Markov Chains.

Results of Markov for MLB and the Astros and Dodgers

For the first model of the entire MLB we have 25 states. We do not need to examine the 25th state as that is when there are three outs and the half inning is over. Let's take a look at five states in particular over the course of the 2017 season and then for the Dodgers and the Astros. We run the tests of the expected runs per state and these results are in our figures. Figure 1 shows the MLB results as whole, figure 2 is the Dodgers, and figure 3 is the Astros. The first state let's look at is 0000 which represents no outs and nobody on base. The expected runs for that half inning is .54 for the MLB. For the Dodgers it is .55 and for the Astros it is .63. Now let's look at the highest expected runs. This is when nobody is out and the bases are loaded. This is represented by 0111. The expected runs is 2.35 for the MLB. For the Dodgers it is 2.17 and for the Astros it is 2.41. The next state we look at is 1101. This is where there is one out and a runner on first and third. The expected runs is 1.22 for the MLB and for the Dodgers it is 1.13. The Astros had an expected run prediction of 1.21. The fourth state we can look at is 2010. This is where we have two outs and a runner on second. The expected runs is .32 for the MLB. For the Dodgers it comes in at .34 and the Astros came in at .38. Finally let's look at the least state where we can expect runs 2000. This is where we have two outs and nobody on. The expected runs is .11 in the MLB. For the Dodgers it comes in at .12 and the Astros comes in at .15. So as expected the most expected runs for this model is when we have the bases loaded with nobody out. The least for all of these models is when there are two outs and nobody on base. We can do a comparison between the Dodgers and Astros in each state not including the 25th state where there are already three outs. The Astros have a higher expected runs in 20 of the 24 states or 83% based off of the 2017 MLB play by play for both teams. This comparison is shown in figure 2 and 3 which highlights which teams has a higher run expectancy at that particular state. Before running our simulation we would predict the Astros would win the game more times than the Dodgers and average more runs.

Markov Chains Prediction for Game 7 of the World Series and Managerial Advice

The Markov Chain is a mathematical model that we can use to help predict results in Major League Baseball. The Markov Chain uses the 25 different states and event data to help predict the runs a team will score. I think Markov Chains can be helpful and I think these models can be helpful to look at. However, I would tell a manager that looking at the team statistics and expected runs from the team at each state, is much more important that looking at individual players and trying to simulate games. The entire year does give us a pretty good sample of how each team performs at the different states. The manager will be able to look at how the team is doing in certain situations. I think the simulation with the individuals is neat, but not as useful because it is a team sport and we had a specific lineup in the simulation. I think part of the limitation is baseball is a team game so looking at an individual only is not that effective. Throughout the year there will be many different lineups and many different players, so looking at team performance as a whole is a better indicator.

As mentioned above we make 16 individual Markov Chains and then put them into the lineup. . The order for the Dodgers is Chris Taylor, Corey Seager, Justin Turner, Cody Bellinger, Yasiel Puig, Joc Pederson, Logan Forsythe, and Austin Barnes. For the Astros the starting lineup is Jose Altuve, Alex Bregman, Carlos Correa, Marwin Gonzalez, Yulleski Gurriel, Brian McCann, Josh Reddick and George Springer. We then simulate the game 10,000 times in order to predict who will likely win game 7 and what the final score will most likely be. After running the results the Astros predicted score was 6.05 and the LA Dodgers had a predicted score of 5.81. Out of the 10,000 simulations the Dodgers were predicted to win 4819 and the Astros were at

5181 meaning the chances the Dodgers would win are at 48%. The actual final score of game 7 of the World Series was the Houston Astros beat the LA Dodgers 5-1. This shows that even though the simulation of the score was not correct, the predicted result of the Astros winning was the same as the actual result and our prediction was correct.