Michael Pallante
MSDS 456
Assignment 3

The purpose of this assignment is to study the events of a baseball game from modeling it, to simulating it, and then predicting it. The game of baseball is one in which analysts can break down an entire game by the events that occur in each half inning of the game. These events each half inning can be modeled as a finite Markov chain, in which there are 24 states of events in a half inning with the 25th state marking the end of the half inning. From there, analysts can use transition probabilities from each state to state to be representative of the events that occur each half inning. Using these transition probabilities, analysts can perform game simulations, and then those game simulations can be used to make predictions of the winner of a baseball game.

With the event data from every game of the entire 2017 MLB season, we are given the ability to build Markov chain models and then use them to estimate transition probabilities, not only for the entire MLB but also for individual teams as well. In this specific case, we will be studying both the Houston Astros and the Los Angeles Dodgers. These two teams faced off in the 2017 World Series in one of the more competitive World Series in recent memory, with the Astros defeating the Dodgers in 7 games to win their first championship in franchise history. To recap, the Houston Astros finished the 2017 regular season with the second best record in the American League with a 101-61 record, and defeated the Boston Red Sox and New York Yankees in the divisional and league championship rounds respectively to reach the World Series. The Los Angeles Dodgers finished the 2017 regular season with the best record in the National League with a 104-58 record, and defeated the Arizona Diamondbacks and Chicago

Cubs in the divisional and league championship rounds respectively to reach the World Series. In the World Series matchup, the Astros won games 2, 3, and 5, while the Dodgers won games 1, 4, and 6, to tie the series 3 games to 3 entering into a decisive game 7, where the winner would be crowned champion. Since the Dodgers had home field advantage for the series, this game 7 was played in Los Angeles and observed the National League rules, where there is no Designated Hitter and the pitchers do bat. The starting pitching matchup for the game was Charlie Morton of the Astros versus Yu Darvish of the Dodgers. The starting lineup for the Astros included Jose Altuve, Alex Bregman, Carlos Correa, Marwin Gonzalez, Yulieski Gurriel, Brian McCann, Josh Reddick, and George Springer. The starting lineup for the Dodgers included Chris Taylor, Corey Seager, Justin Turner, Cody Bellinger, Yasiel Puig, Joc Pederson, Logan Forsythe, and Austin Barnes. Using estimated transition probabilities for the Houston Astros and Los Angeles Dodgers, we can guide game simulations in which we can build a game simulation model that predicts the winner of that decisive Game 7 of the 2017 World Series between these two teams. We already know the Houston Astros defeated the Los Angeles Dodgers 5-1 to win the World Series, however, it will be interesting to see if the game simulation model yields different results for the Game 7 winner.

Transition probabilities were not only estimated for the MLB, Houston Astros, and Los Angeles Dodgers, but they were also estimated for the position players for the starting lineups of the Astros and Dodgers rosters. To do so, we set the starting lineups of both teams in the exact order in which they were listed in the previous paragraph. For the Markov chain transition probabilities, we will be using 10,000 simulation runs to estimate end of inning runs scored associated with each of the 24 half inning states. Since there are 8 position players in each team’s starting lineup for a total of 16 starting position players, a total of 16 Markov chain models will be needed, or 1 for each starting position player. These 16 Markov chain models will be able to inform our game simulation model for the 2017 World Series Game 7 between the two teams. Pitchers were ignored in our Markov chain models, as batting data for pitchers is severely limited. So, in our models, we created the assumption that the pitchers will always strike out. In summary, at the conclusion of this process, we will run our game simulation 10,000 times to predict the probability that the Los Angeles Dodgers and Houston Astros win the 2017 World Series Game 7.

Now, we will take a deeper dive into the step by step process that was used to achieve the goals of this assignment. First, we obtained the 2017 MLB regular season game data and preprocessed the event data so that we could view the game data of every MLB team. This yielded csv files of game data for all 30 MLB teams and these files were then read into R and further organized using column names and indicators for runners at each base. A variable was then created to indicate the starting state of each play and every state that follows. States were organized so that it read outs, runners on first, runners on second, and runners on third, in that order. For example, the state 1010 represents one out and a runner on second, the state 2101 represents two outs and runners on first and third, and the state 3000 indicates that three outs have been reached and the half inning is over. Next, two-way frequency tables of transition states were created. These tables were then converted into probability transition matrices and then into a Markov chain.

Continuing on with the MLB data, we calculated the expected runs scored moving from one state to another and then looped through all possible states to calculate how many runs were expected to score by the end of an inning, given an initial state. This concluded our work with the full MLB data and we then shifted our attention to the Houston Astros and Los Angeles Dodgers teams. All that was done here was that the MLB data was subset for the Astros and Dodgers only. The exact same step by step process as above was carried out for the Astros game data and the Dodgers game data. From this, we gained the expected runs for both teams.

Once we finished with the subset data for the Astros and Dodgers and got their expected runs, we transitioned into the building of transition matrices for each individual starting position player on both teams. As mentioned before, this will give us a total of 16 starting position player matrices. We followed this with the creation of a transition matrix for the pitchers, which as stated before, will be ignored. So, the transition matrix was built to assume that the pitchers will always strike out. After these matrices were built, we started the game simulation process.

To reiterate, our goal was to simulate the 2017 World Series Game 7 between the Houston Astros and Los Angeles Dodgers. The actual outcome in real-life had the Astros winning 5-1 to win the game and the series, becoming the MLB champions. Using our models, we simulated 10,000 games between the two teams to see who was more probable to win Game 7. Our simulations agreed with the real-life result and had the Astros winning Game 7 in more simulations than the Dodgers. The Houston Astros won 5,290 of our game simulations, while the Dodgers won 4,709 of our game simulations. This means the Astros won approximately 53% of our simulations, while the Dodgers won approximately 47% of our game simulations. The Astros also averaged approximately 6.1 runs scored in our simulations, while the Dodges averaged approximately 5.77 runs scored. To visualize these final results, they can be seen in Figure 1.0, Figure 2.0, and Figure 3.0 below.

Figure 1.0

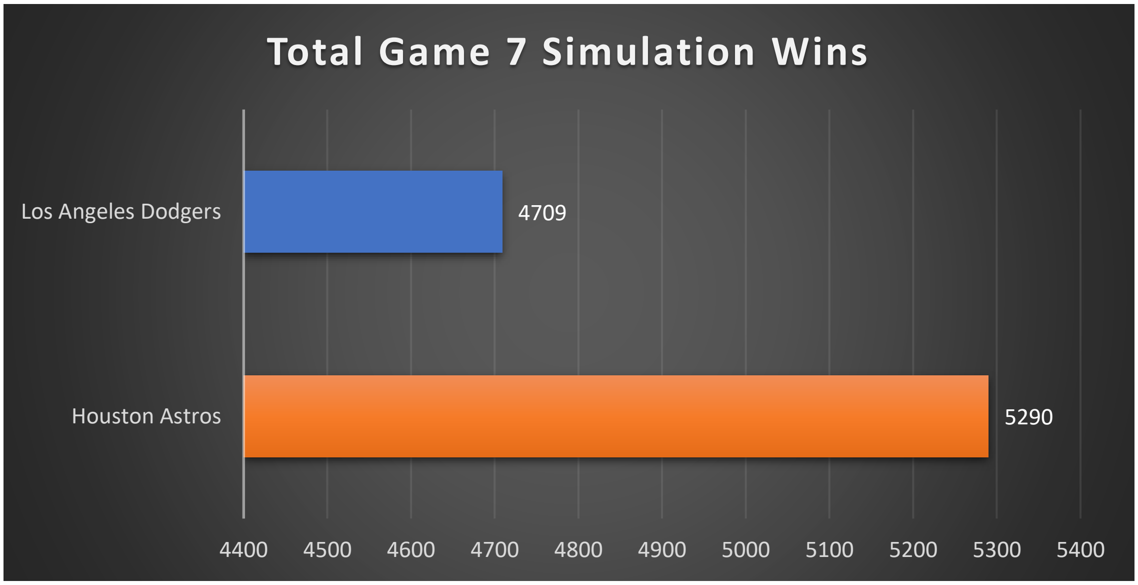


Figure 2.0

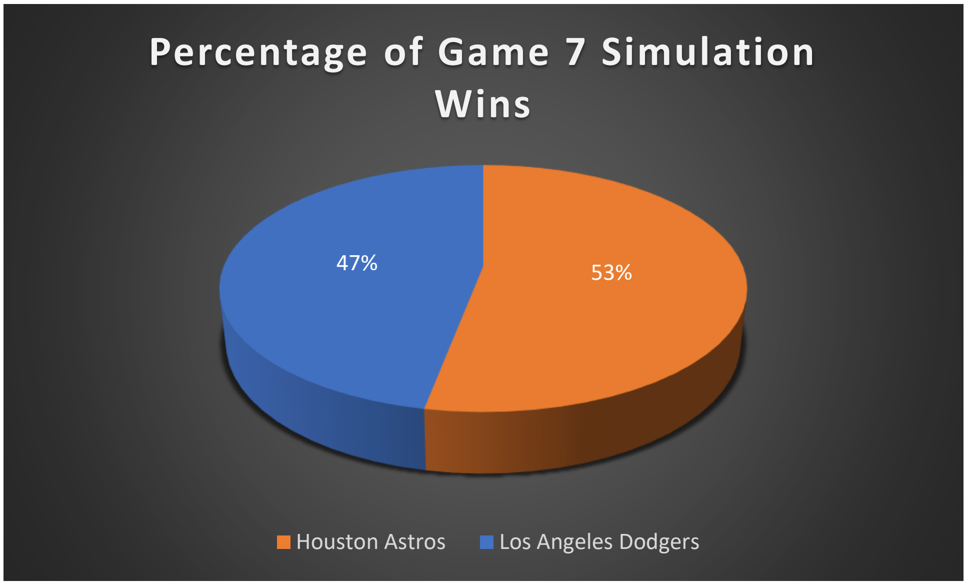
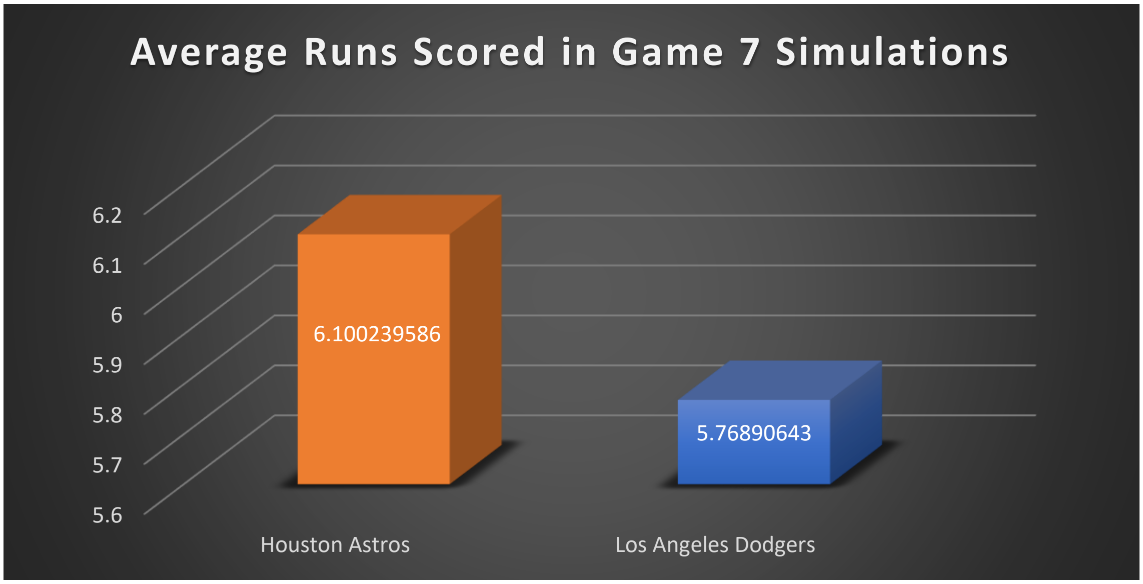


Figure 3.0



The key takeaway of this assignment is seeing the predictive power of the game simulation models employed. While no simulation can be perfectly accurate and may not be completely predictive of the results of a game, the completion of so many simulations does give an idea of the game results we should expect. In the case of the teams involved here, managers and decision makers for the Houston Astros and Los Angeles Dodgers can build such models and tinker with their lineups to see what gives the best chance of yielding a win. There is a lot of opportunities in game by game lineup construction for baseball teams using these models. There are definitely some difficulties in employing Markov chain models at the individual player level though and teams must be aware of that. As always, it is not smart for teams to use this method as the one and only method of predicting a win in a game, but rather a good idea to use to supplement all other tools and resources they have at their disposal.