Data Analysis on Baseball Teams and Factors that Increase the Attendance

When examining which league--American League (AL) or National League (NL)--in Major League Baseball would be better suited for a new team to start and win more games, both leagues have show similar patterns of winning. No one league seems to have dominated another. Rather, they were pretty much identical when looking at the PDF and CDF of wins for both leagues. However, the PDF plot of wins for AL and NL has revealed a slight difference between them. First, NL has shown more teams to have won between 80 and 90 games than AL. In baseball, winning 80 to 90 games can be considered competitive and playoff worthy for many teams. However, between 90 and 100 plus games, AL seems to have more teams that have won. Typically, a team that has won more than 90 usually clinches a spot in the playoffs. As for 100 plus game winning teams have shown to have won either its division or pennant during a regular season even though winning more than 90 games does not always guarantee an ultimate prize in baseball—A Word Series title.

Also, lately baseball has lost many fans due to a slow pace and nature of the game. Some of the features of the game have been said to engender more actions around the bases (or diamond). Namely, doubles and triples cause more excitement and actions by baserunners. Thus, I wanted to find out if there’s a correlation between doubles and attendance for the games. However, based on the outcome of the correlation analysis, there didn’t seem to be a strong relationship. I also looked into the effect of single hits on attendance. But again, there was a weak correlation between them.

I think the correlation analysis between action-inducing features of the game and attendance may yield different results if I resample the data or reduce the scope of the data. Even though, I removed the rows with zero attendance data, the number of games played varied over the years. It wasn’t until 1980s that MLB started a 162-game season. In the early 1900s, there were some years when fewer games were played perhaps due to wars or economic depressions. Thus, a smaller population could have produced a different outcome on the correlation analysis.

One other variable that I did not examine this time was stolen bases. Stealing bases by runners naturally brings more actions by the players and excitement from the fans. Thus, a new correlation analysis and a hypothesis test could affirm or deny a theory by MLB that more actions around the bases will bring back more fans to the games.

One of the assumptions that I made was incorrect. For example, strikeouts by pitchers (SOA) would be exciting for the fans since that means a great pitching duel. But it seems to have a slightly negative impact on the attendance since an increase in SOA could cause a decrease in attendance as revealed by my regression analysis and its small negative intercept number. And that in a way makes, though. As more strikeouts by pitchers increase, less actions around the bases, leading the fans to believe that the game is less exciting, less scores, and boring in some cases. Thus, the attendance could decline as a result. Also, I assumed that American League has dominated the competition due to designated hitter rule where a hitter rather a pitcher hits for the 9th spot of the batting order. But both AL and NL have fared similarly in terms of wins.

Lastly, I wanted to perform a correlation analysis between doubles/triples and triples and plot normal probability using the Lognormal Distribution. However, I kept encountering errors related to computation of log on doubles and triples. I could not get my code to work and plot the normal probability using the Lognormal Distribution.