Statistical Analysis Behind Signing All-Star Players

Emily Hou, Tony Yao

Boston University

CS591 C1 Computational Tools for Data Science

Abstract

With Major League Baseball's offseason starting up, we plan on determining whether or not it's worth it for baseball teams to sign free agents to large contracts. We will look at all-star game appearances and analyze whether the higher paid players are making the all-star team more than the lower paid players. Our expected results are that over the last 30 years, higher paid players have not been making the all-star game significantly more than the lower paid players. We plan on performing linear regression and time series to analyze whether or not the best performing baseball players are actually the ones who are being paid the most. We will value players who make the all-star game as the players who perform the best that year. By analyzing the results, we hope to find that having a higher salary isn't an indicator of making the all-star game more often.

Introduction

Every year around this time, 30 Major League Baseball front offices work tirelessly to improve their ball club. They have to decide whether it's worth it to sign star players to large contracts or try developing their young guns. Due to how the baseball salary system works, young players have a rookie contract of 6 years with their team when they first make it to the majors, so they don't end up getting their big payday until they're in their late 20's or early 30's. Teams need to decide whether it's worth it to give long and expensive contracts to these players who are still barely in their prime, knowing that the later years would be paying them when they're in decline. An additional factor that the teams will need to consider is whether the player's quality of play will drop due to complacency, knowing that they already got their big payday. Therefore, front offices must decide whether or not to sign these well known free agents to long and expensive contracts with hopes of their past performances continuing.

The datasets that we plan to use are from the Baseball Databank, which is a compilation of historical baseball data distributed under Open Data terms. The database contains complete batting and pitching statistics from 1871 to 2015, fielding statistics, team standings, team stats, managerial records, post-season data, and more. It is a collaborative collection on player performance from the Lahman Baseball Database, identification and demographics from Chadwick Baseball Bureau, and game logs published by Retrosheet. The datasets are available on github, https://github.com/chadwickbureau/baseballdatabank, and Sean Lahmna's website archive, http://seanlahman.com/baseball-archive/statistics/. Because the datasets are in csv format, stored within a zip file, pre-processing of the data is not needed. We plan on using this dataset to analyze the relationship between players' salaries and the number of all-star game appearances they make.

Datasets

The three main datasets we used are:

Salaries.csv - includes individual players' salaries by playerID, along with information on the team that they are from and the year that they played on that team.

AllstarFull.csv - includes which players made it to the all-star game each year by playerID. We plan to use this dataset to analyze whether the players who made the all-star games were the ones who were highly paid.

Master.csv - includes information for each player by playerID, along with personal information outlining age, first/last name, debut date, most recent game played, and birth/death dates. Because we are only interested in players who are still alive, we'll be doing a simple trim on the data, as well as associate playerID with the players' actual full names.

From the datasets, we cleaned the data to create three new datasets (see getDatasets.py) and we used the following merges:

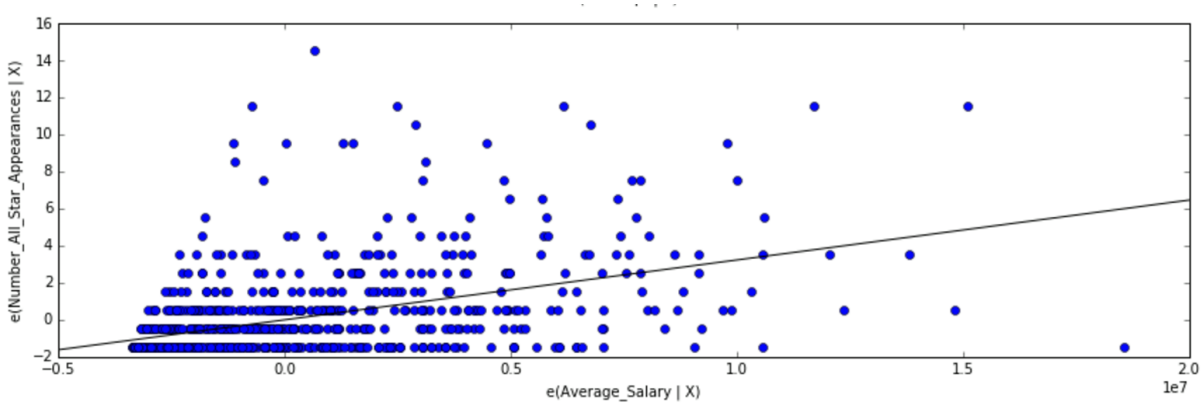
PlayerAll: Merge AllStarData.csv with MasterData.csv with SalariesData.csv by playerID to get a dataset with the year, team of player, full name of player, age of player that year, salary of player that year, and whether the player made it to the all-star game that year.

Techniques

In this project, we plan on implementing two major techniques:

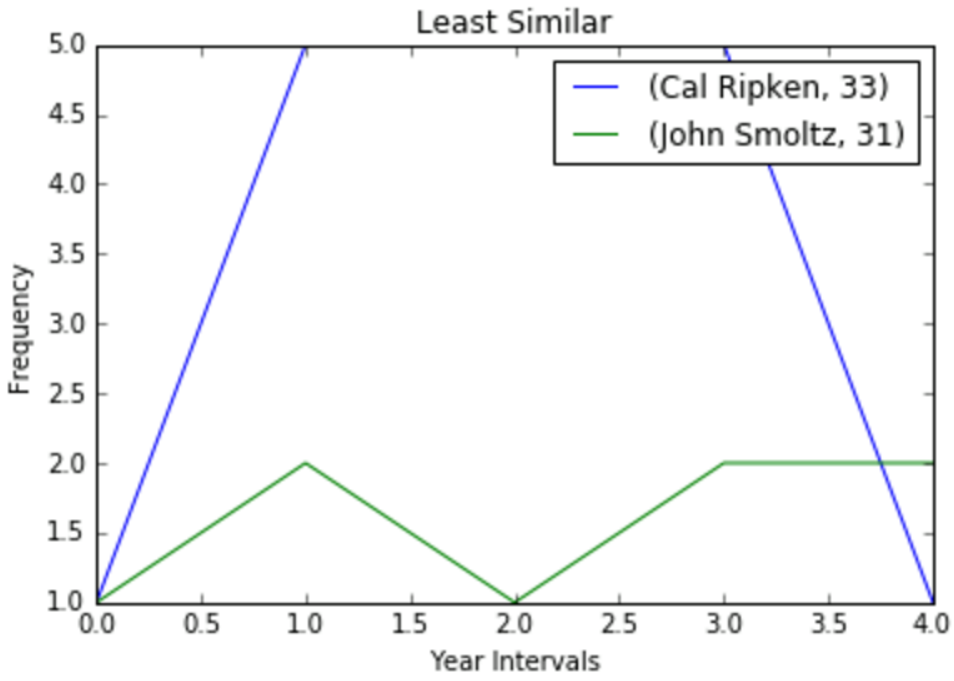
Linear Regression:

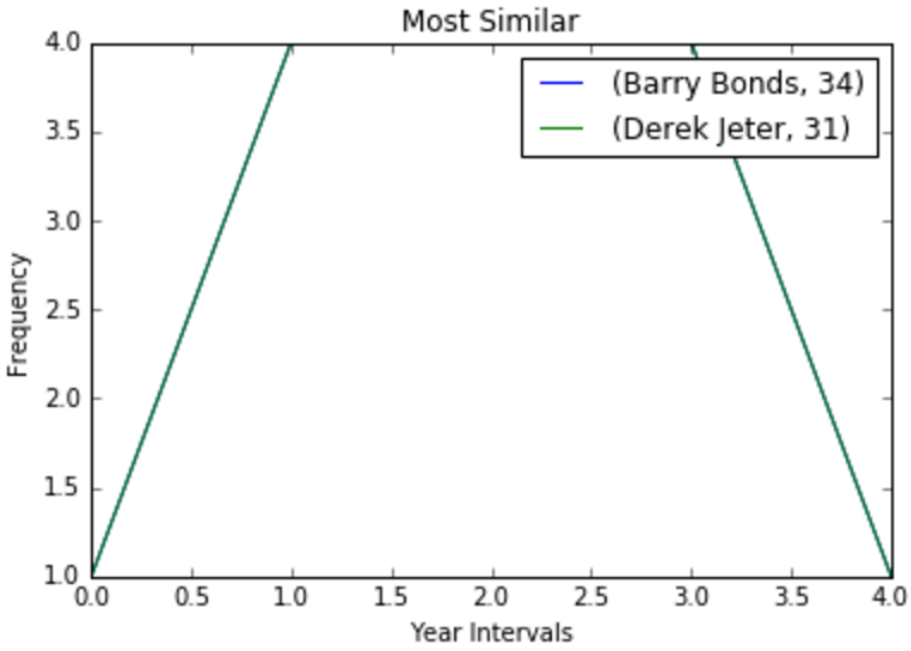
Our model of linear regression measures average salaries of individual players against the number of all-star game appearances for those players. In this instance, the independent variable is collected as their average salary (total salaries throughout the years that the players have played over the total number of years) and the dependent variable is the total number of all-star game appearances for each player. We analyzed the R-squared value of 0.219 to determine that the linear model does not provide a well-defined linear regression line as a predictor for future all-star game appearances. The small R-squared value suggests that players who earn more money on average are not guaranteed to have more all-star game appearances. This is due to the weak correlation between the two variables.



Time Series:

Since the data is over the span of 30 years, we utilized time series to track down total all-star game appearances in 5-year intervals. We analyzed the top 100 players with the most all-star game appearances and plotted it by intervals after using Minkowski distance as our metric. Because Minkowski is within a normed vector space that could be considered a generalization of both the Euclidean and Manhattan/Hamming distance metric, we thought it would be interesting to see what results this "mixed" metric may deliver and analyze its results. Minkowski, by definition, relies on the order of the p-value between two point to calculate distance. We found the two least similar and most similar players in their trends of all-star game appearances in our sampling to gain insight into potential similar patterns that may exist for the future, as stated in our original hypothesis.





Results and Discussion

From the regression model, we see that the R-squared value was not very high with a value of 0.219, which illustrates that the linear model isn't a good predictor for future all-star game

appearances for each player. The number of future all-star game appearances for a particular player isn't guaranteed to increase or decrease by much depending on the trend of his past performance. Therefore, the results support that having a higher average salary isn't a good predictor of reaching more all-star games. Overall, it seems that the model applied was not as statistically significant as we had thought.

From the time series, we found that Cal Ripken and John Smoltz had the least similar performance trends, focusing on their respective intervals. Ripken ended up making a lot more all-star games than Smoltz at an older age, which wasn't a surprise since Ripken was nicknamed "The Iron Man" due to playing 2,632 games consecutively and making 19 all-star games in a row into his 40's. On the other hand, for the most similar, Barry Bonds and Derek Jeter had the same exact trend, despite the fact that Bonds started his career nearly a decade before Jeter. Although Bonds and Jeter had the same trend, their average age of making all-star games differed since Bond's average age was 34 years old, while Jeter's average age was 31 years old. This shows that Bonds continued playing at a high level past his prime, but it was later discovered that Bonds used human growth hormones for recovery in his later years. Upon further research, Jeter made all of his all-star game appearances with his same original team, while Bonds was later signed by a different team to the largest contract in baseball at the time.

Conclusion

We expected to find that players with higher salaries didn't indicate that they would make more all-star games. From looking at the linear regression, we found that there is a weak correlation between a player's average salary and their all-star game appearances. Additionally, our time series indicated that the two most similar players in terms of all-star game appearances in 5-year intervals were Barry Bonds and Derek Jeter. Although both of these players are legends in their own right, their careers tell different stories. Derek Jeter was the young homegrown talent that stayed with his original team his whole career, while Barry Bonds was the star free agent who signed an expensive contract. Since both players' all-star game appearances were so similar, it is further proof that you don't need to sign expensive free agents to long contracts when you could trust in your homegrown talents.

Instead of signing these aging players to expensive contracts based on their past performances, baseball front offices should look for cheaper pieces to build around their homegrown core. Just by looking at the past couple of world series winners: the Chicago Cubs, Kansas City Royals, and San Francisco Giants were all teams that did not have the highest payrolls, but were instead comprised of mostly homegrown talent. Meanwhile, many of the teams with the highest payrolls didn't even make the playoffs, due to being stuck with stars that were in decline. Therefore, baseball front offices should not sign these well known free agents to long and expensive contracts, but instead invest in developing their farm system, so that they will have young players on inexpensive rookie contracts that will one day make up the core of their championship winning team.

Future considerations of our project could focus on analyzing any correlations between player injuries and their respective age; for example, number or severity of injuries over the course of their playing career. This could provide another perspective in whether or not front offices should sign aging star players. Additionally, iterations of this project idea could further explore whether the weak correlation between all-star appearances and average salaries could be translated on a team-based scale. That is, whether lower team payrolls experience a similar success. Since the Cubs, Royals, and Giants were able to find success without many highly paid stars, it would be interesting to see whether this trend has been evident over the past 30 years.