Statistical Analysis Behind Signing Star Baseball Players

Emily Hou and Tony Yao
CS505 Computational Tools for Data Science
Boston University

ComputerScience

Introduction

During this time of the year, major league baseball front offices have to decide which players they want to target for their team to sign. Since baseball players have a rookie contract of 6 years with their team when they first make it to the majors, most of them enter free agency in their late 20's or early 30's. Although players will still be in their prime, signing them to an expensive yet long contract would mean paying them when they're in decline. Therefore, front offices must decide whether or not to sign these well known free agents to long and expensive contracts, in hopes of their past performances continuing.

Objective

We plan on performing two methods 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.

1. Linear Regression between all-star game appearances and average salary.
2. Time Series of players who make the all-star game in 5-year intervals over the last 30 years.

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.

Datasets

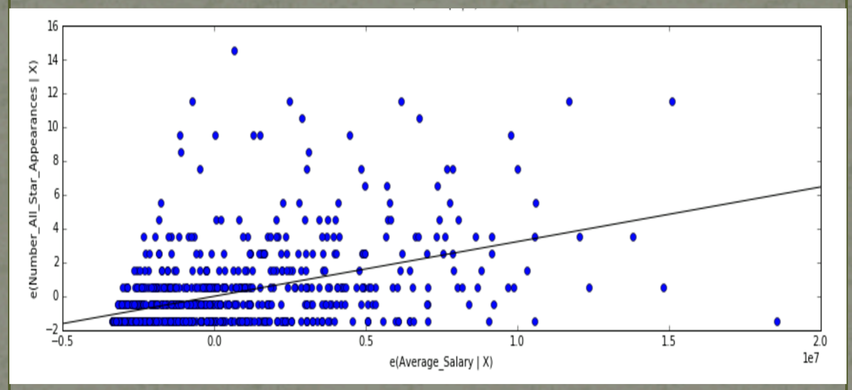
The datasets we used are from Baseball Databank.

1. Salaries.csv : Includes players' salaries and the teams they played for that year.
2. AllstarFull.csv: Includes which players made it to the all-star game each year.
3. Master.csv: Includes each player's background information such as their age, first/last name, debut date, and most recent game played.

We merged the three datasets to form PlayerAll.csv, which includes the year, team of player, full name of player, age of player that year, salary of player that year, and whether they made it to the all-star game.

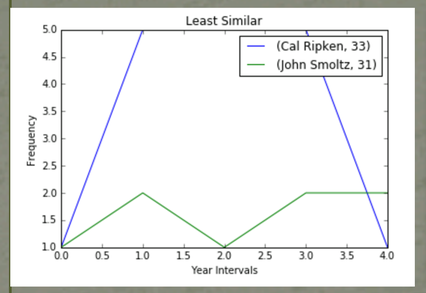
Method 1: Linear Regression

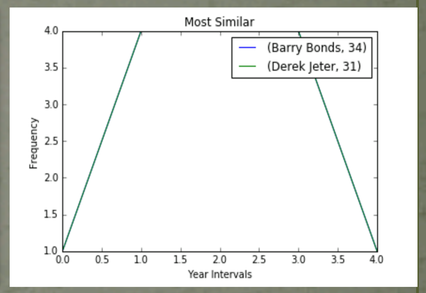
Our model of linear regression measures average salaries of individual players against the number of all-star game appearances for those players. In this instance, the independent variable is collected as their average salary (total salaries throughout the years that the players have played over the 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.



Method 2: Time Series

Minkowski Distance





Since the data is over the span of 30 years, we utilized time series to track down total all-star games 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. We found the two most similar and least 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. We found that Ripken and Smoltz had the least similar performance trends, focusing on their respective intervals. On the other hand, for the most similar, Bonds and Jeter had the same exact trend, despite the fact that Bonds started a few years before Jeter. Upon further research, Jeter made 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 will 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 who 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.