

**Analysis of Baseball Databank**

***Danny Harris,*** *IU School of Informatics and Computing, IUPUI*

***Matthew McGee,*** *IU School of Informatics and Computing, IUPUI*

***Antonio Fernandez,*** *IU School of Informatics and Computing, IUPUI*

***Chris Cousert,*** *IU School of Informatics and Computing, IUPUI*

***Abstract --*** *The dataset is used to determine how the skill of baseball players has evolved over time. This is done using linear regression, based on specific statistics for three different categories--batting, pitching and fielding. Based on findings, there is evidence to suggest that baseball player are getting better over time.*

**Table of Contents**

1. Introduction 3
2. Supporting Research 3
3. Data Analysis 3-5
4. Explanation of Data 3
5. Attributes 4
6. Procedure for Analysis 4
7. Data Summary 4
8. Conclusion 4-5

**IV. Linear Model Examples 5**

V. References 5

***Keywords -- Baseball; MLB; Regression; Linear Regression***

1. **Introduction**

The overall objective of this project is to compare different sets of data from major league baseball players over several decades to see if we find any trends or interesting findings based on our analysis. The null hypothesis is that the average skill level of baseball players has stayed the same over time. To determine this, we observed what stats are correlated to a players’ success and those specific thresholds. Doing this allowed for a better understanding of how the skill level of the players have changed overtime whether it be for the better or worse.

By analyzing and visualizing the dataset, the metrics were then separated into 3 categories -- batting, pitching and fielding. At this point, a determination was made of which attributes were going to be used from each of these categories. The idea was to choose attributes that would be most representative of what it means to be a quality baseball player, regardless of the time a given player was active.

These parameters included statistics such as home-runs, hits, strikeouts, walks, assists, and fielding errors. Upon the completion of this process we felt confident in continuing forward into the research stage of this project. Utilizing many data analysis techniques in the programming language R, we were able to compile multiple graphs that showed trends in these statistics throughout the years.

1. **Supporting Research**
2. *Hitting is Contagious in Baseball: Evidence From Long Hitting Streaks by Joe R. Bock*

This study dives into the legend of the “hot hitter” using data analysis. The study tested whether or not a hitter that has a consecutive game hitting streak can affect their teammates ability to get on base. The study concludes that there is evidence to suggest the potential existence of a “statistical contagion effect.” This was a very interesting study and it also gave us a good idea of what kinds of attributes would be good to use in our model, for batting.

1. **Data Analysis**
2. Explanation of Data

Our data resource is an open source dataset from kaggle.com called Baseball Databank. This data set includes twenty CSV files of data from baseball teams, including both players and managers. We very quickly whittled this dataset down to a much more manageable three CSV files. These three CSV files were the data for batting, hitting and fielding. Before consolidation, there were instances of numerical (ratio and interval), categorical (nominal and ordinal) data, and data types ranging from strings, to floats, to integers. Our study is broken into three parts, as stated above. For the batting portion of the study there was a focus put on three attributes--homeruns, hits and RBI. For the pitching portion there was a focus put on three attributes--strikeouts, ERA and walks. For the fielding portion of the study there was focus on three attributes--errors, assists and double plays.

1. Attributes

Originally our dataset contained over 20 csv files. It contains data from seasons in 1847 to the year 2015. This data set includes 20 CSV files of data from baseball teams, including both players and managers. Each player has a unique ID that links to the master table, most of the 20 tables range from stats (batting, fielding, pitching), to salaries, to awards, to team rankings, as well as other statistics. Most of the tables have 7 or more columns of data, with the largest having 48. There are instances of numerical (ratio and interval), categorical (nominal and ordinal) data, and data types ranging from strings, to floats, to integers. The specific attributes we decided to analyze included: Home-Runs, Hits, RBI’s (Runs Batted In), Strikeouts, ERA (Earned Run Average), Walks, Errors, Assists and Double Plays. These attributes are good indicators of a players’ skill level.

1. Procedure for Analysis

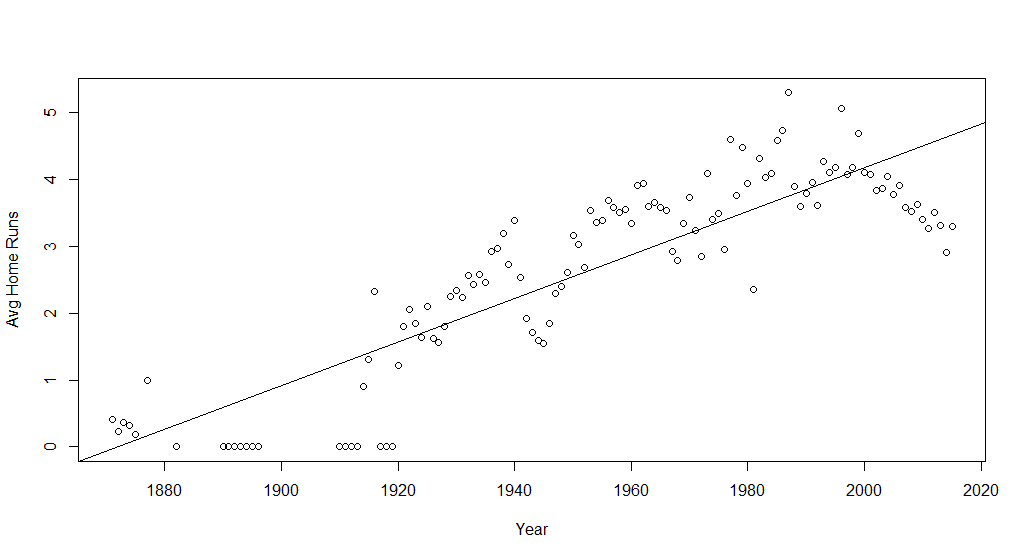
First, the data that was not relevant to our study was removed. For example, there were tables that pertain to team managers and player salaries. These tables were dropped, as they are not relevant. The next step was to remove duplicate or incomplete observations from the dataset. Next, any inconsistencies in class labeling conventions were identified and corrected. For example, referring to the “American League” in one observation and “AL” in another observation. It seems that in the earlier years of baseball, statistics weren’t kept as rigorously as they are now. So, there is some missing data that we certainly had to deal with in this respect. For example, it doesn’t seem as if “intentional base on balls” or intentional walks, was a statistic that was kept early on.

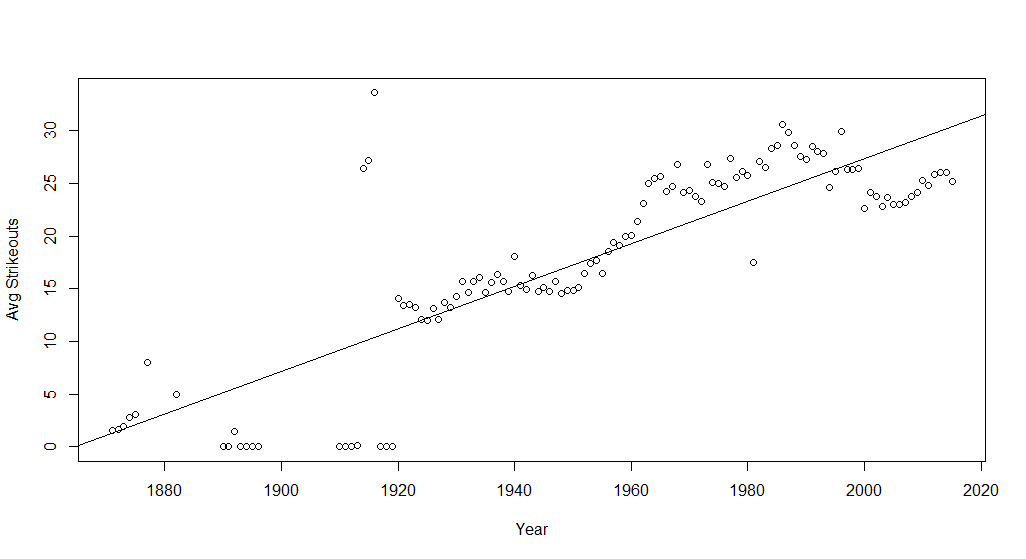
After choosing the metrics data cleaning was performed, followed by the calculation of averages for each of the given statistics and subsequent grouping by year. Averages were chosen for each metric to offset an observed fluctuation and general increase in the number of players by year. Because of the sheer size of the dataset, any rows with missing values were dropped before the calculation of averages. Furthermore, filling these null values with an average would hinder the ability to compare averages across time, as it would make it difficult to determine how much the overall average was affecting the yearly average.

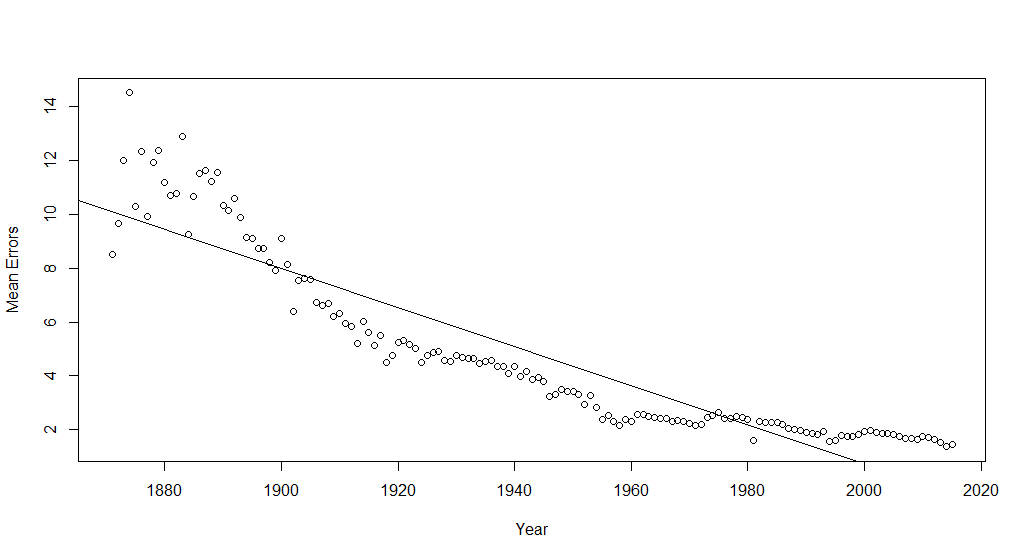
D. Data Summary

Linear models showed that for nearly all of the chosen positive statistics (home runs and RBIs for batters and Strikeouts for pitchers) increased over time, while most negative statistics (errors for fielders and walks for pitchers) decreased. Hits and RBI’s (batters), double plays (fielders), and ERA (pitchers) remained relatively stable after some increase up until around the 1900s. The only statistical trends that identified any worsening play ability were increased strikeouts(batters) and decreased assists (fielders). It should be noted, however, that assists have remained relatively stable since 1960. Also, we must consider that both pitchers and batters are inversely involved in the strikeout statistic, meaning that aside from stability, one of these groups increasing in trends will result in a decrease of the other group.

**Linear Model Average Examples**

**Figure 1: Home Runs**

**Figure 2: Strikeouts**

**Figure 3: Errors**

F. Conclusion

Given that the majority of the metrics point to the improvement of players all around, it can be reasonably concluded that baseball players have gotten better over time, meaning a rejection of the null hypothesis and support for the alternative hypothesis.

It should also be noted that jumps, dips or spikes in many metrics (assists, double plays, and others) appear during or around the time of WWI and/or WWII. In 2001, the year of the 9-11 tragedy, significant jumps (increase) were also observed in double plays and assists. Documented possible explanations include players leaving to serve in the World Wars and changes to scheduling in 2001.

**IV. References**

**[1]** Sports, O. (2019, November 17). Baseball databank. Retrieved May 05, 2021, from [Baseball Databank](https://www.kaggle.com/open-source-sports/baseball-databank)

**[2]** Bock, J. R. (2012). Hitting is contagious in baseball: Evidence from long hitting streaks. PLoS ONE, 7(12). doi:10.1371/journal.pone.0051367

**[3]** Sherman, J. (2001, February 2). Retrieved May 07, 2021, With Unbalanced Schedule, It’s A Brand-new Ballgame; Changes Could Benefit Yanks, Mets, <https://nypost.com/2001/02/02/with-unbalanced-schedule-its-a-brand-new-ballgame-changes-could-benefit-yanks-mets/>

**[4]** Imagine Sports Editorial Team. Diamond Mind Online. Retrieved May 07, 2021, What Was Baseball Like During World War 2?, https://imaginesports.com/news/baseball-during-world-war-2