# Data Science 2: Statistics for Data Science

## Report on the Analysis & Predictive Modelling of Strikeouts vs Home Runs in Major League Baseball Players 1871 - 2020

### Submitted by Group 8

Han Chen

Andrew Lagree

Rebecca Manning

Charlie Wang

Nathan Ye

Johanna Yu

### Supervised By:

Dr. Sergiy Nokhrin

Dr. Amanda Easson

Dr. Prakriti Sodhi

December 2021

# Objectives

What are you setting out to prove or predict? What is your rationale for there being a correlation in the data that you’re looking to confirm and/or exploit?

The objective for this project is to investigate 1) the relationship between Strikeouts (SO) and Homeruns (HR) in the history of baseball, and 2) if the relationship (ratio?) of SO and HR varies among the eight different eras of baseball. In order to hit a homerun, the player will swing the bat and swinging the bat will create opportunities for strikeouts. We believe that there will be a positive relationship between SO and HR. We also noted on the changes in baseball rules in the eight eras and these rules may have a confounding impact on the relationship (ratio?) of SO and HR.

# Data Preparation

We used an online baseball statistics database from <http://www.seanlahman.com/baseball-archive/statistics/>. The databases are free, with copyright 1996-2021 by Sean Lahman. The website contains complete pitching, batting, and fielding statistics from Major League Baseball from 1871 to 2020. It includes data from the two current leagues (American and National), the four other "major" leagues (American Association, Union Association, Players League, and Federal League), and the National Association of 1871-1875).

For our assignment, we used the batting statistics database (Batting.csv). There were 22 column variables, with 108,789 row entries. The column variables were defined as follows:

playerID Player ID code

yearID Year

stint Player's stint (order of appearances within a season)

teamID Team

lgID League

G Games

AB At Bats

R Runs

H Hits

2B Doubles

3B Triples

HR Homeruns

RBI Runs Batted In

SB Stolen Bases

CS Caught Stealing

BB Base on Balls

SO Strikeouts

IBB Intentional walks

HBP Hit by pitch

SH Sacrifice hits

SF Sacrifice flies

GIDP Grounded into double plays

For our assignment, we focused on data for 4 variables of interest, specifically playerID, yearID, HR, and SO.

Missing Values:

Overall the data quality was very good. However, there were 2,100 missing values for the SO variable (comprising 1.9% of the total 108,789 row entries). There were a few ways we could have dealt with those 2,100 missing values for SO:

a) We could fill in the missing values with the median or mean of the SO variable

b) We could impute a value based on grouping of similar characteristics for other variables

c) We could drop the rows with missing SO values

Because we wanted to have the most accurate estimate of the association between HR and SO, we wanted to minimize bias and confounding. We felt that trying to fill in missing values with the median or mean, or trying to impute values could have created some bias and confounding. We wanted to have the purest unaltered data available.

Therefore, we felt that it was most appropriate to drop the rows with missing SO values. 2,100 missing values out of a total of 108,789 total entries represented only 1.9% of the total sample. Dropping those rows still left 106,689 rows of data. Therefore, dropping the rows with missing values should have a negligible effect on the analyses.

Feature Engineering:

Since we wanted to explore the relationship between HR and SO over the years, we created different bins based on what era of baseball a player played in.

Based on this article from The Sport Journal (<https://thesportjournal.org/article/examining-perceptions-of-baseballs-eras/>), the different eras of baseball can be broken down generally as follows:

Era1 Pre Dead Ball Era (1870-1900)

Era2 Dead Ball Era (1901-1919)

Era3 Live Ball Era (1920-1941)

Era4 Integration Era (1942-1960)

Era5 Expansion Era (1961-1976)

Era6 Free Agency Era (1977-1993)

Era7 Long Ball/Steroid Era (1994-2005)

Era8 Post Steroid Era (2006-2020)

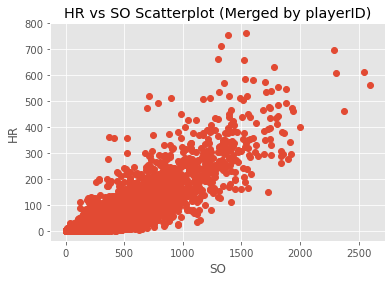
Unique Players:

Inspecting the data frame further, we noted that the same playerID’s appeared in multiple rows. This represented the same player playing in multiple different seasons. After dropping the rows with missing SO values, there were 19,445 unique players among the 106,689 remaining row entries.

Since each observation in an analysis should be an independent observation, when conducting our analyses, we merged the data for rows with the same playerID (e.g., the same player but data for a different season).

# Analysis or model

If you are conducting an inference test, explain the analysis you performed clearly and include well-labelled diagrams to make your points. If you chose to do a predictive model, explain the model, how you trained and tested it, and how well it works. How did you confirm that the data met the requirements for the test or modelling technique to be valid?

Having categorized players into their respective eras and combining the years of each unique playerID, we are able to visualize how many HRs and SOs each player has accumulated throughout the history of baseball. The SO to HR scatterplot shows that many players are closer to the bottom than to the players with the most HRs and SOs. The dataset is positively skewed as 62.7% of players have hit 0 HRs and 21.8% of players have 0 SOs. The 75th percentile of players hit 5 HRs and 105 SOs compared to baseball’s leaders of 762 HRs and 2597 SOs

A trend appears after taking a look at the mean player’s data for a given era or year. Players are striking out more as we progress to more recent history but they are also having more success hitting HRs and thus lowering the amount of SOs for every HR they hit.

Chart, line chart

Description automatically generatedChart, histogram

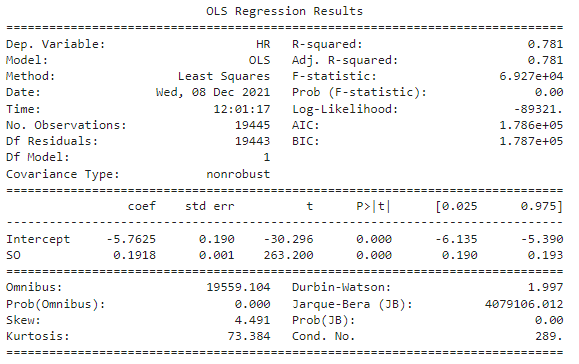
Description automatically generated

The description table of the dataframe shows the above relationship is true across the board. The ratio of SOs to HRs is smaller as baseball has progressed through the eras for the 25th, 50th, 75th percentile.

A picture containing text

Description automatically generated

An OLS model of the overall relationship between HRs and SOs show there is a strong relationship between the two with a R-squared of 0.781. The model predicts a SO:HR ratio of 5.2. The errors are homoscedastic, there is no autocorrelation, the error terms are normally distributed so an OLS model is feasible to use. The addition of the era variable was considered not significant so not included into the overall OLS model.

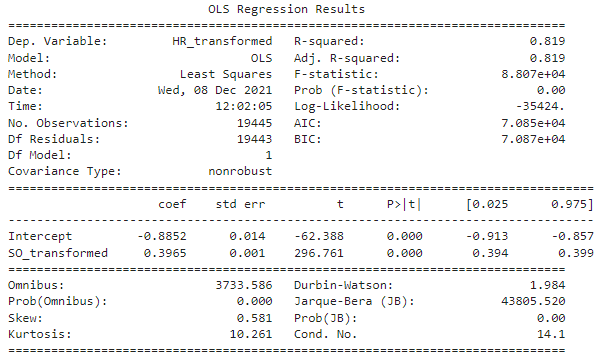


Another OLS analysis was done except this time outliers were removed. 272 outliers were removed but a similar relationship was found. 5.9 SOs were expected for every HR hit in this model. The ratio is higher because the outliers are the players that exceeded in having longer careers and thus big HR and SO numbers. This can be confirmed because it is known majority of the players hit 0 HRs so the overall distribution is positively skewed. With the outliers removed, the skewness is closer to being normally distributed than with the outliers included.

Table

Description automatically generated

A final overall OLS analysis was done by transforming both the SOs and HRs. Both variables were square rooted. This revealed a stronger correlation with a R-squared value of 0.819. This transformation normalizes the dataset. The equation would look like HR0.5=0.3965\*SO0.5-0.8852. The true y-intercept would be 0 since it is impossible to have negative HRs or negative SOs. After back transformation it would look like HR=0.1572\*SO. This value makes sense to be lower than the 0.1692 in the OLS with outliers removed or 0.1918 in the overall OLS since the square root transformation attempts to compress the higher values, so lower values become more spread out.



Lastly, we look at OLS models for individual eras. As the eras progressed, we see a stronger relationship between SOs and HRs starting with a R-squared value of 0.565 in era 1 and culminating with a R-squared value of 0.860 in era 8. The OLS models correctly predicts eras 1 and 2 would have the worst SO:HR ratios but incorrectly predicts that eras 3 and 4 would have the best SO:HR ratios.

|  |  |  |
| --- | --- | --- |
| Era | R-Squared | SO Coefficient |
| Era 1 | 0.565 | 0.1021 |
| Era 2 | 0.563 | 0.0596 |
| Era 3 | 0.623 | 0.2435 |
| Era 4 | 0.706 | 0.2503 |
| Era 5 | 0.768` | 0.1969 |
| Era 6 | 0.810 | 0.1895 |
| Era 7 | 0.817 | 0.2112 |
| Era 8 | 0.860 | 0.1700 |

# Conclusions

Did you prove/disprove your hypothesis or create a useful model? What did you learn about your data set?