# Data Science 2: Statistics for Data Science

## Report on the Analysis and Predictive Modeling of

## Home Runs versus Strikeouts in Baseball Players

## from 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

The objective for this project is to investigate: 1) The relationship between home runs (HR) and strikeouts (SO) in the history of baseball (specifically if SO can be used to predict HR), and 2) If the relationship between HR and SO varies among the eight different eras of baseball.

In order to hit a home run, a player will need to swing the bat. Swinging the bat creates more chances for ball contact that results in HR’s but potentially also more chances to SO. The rationale for the correlation between HR and SO is that players who swing the bat more often will be more likely to hit home runs, but will also be more likely to strikeout. Therefore, we believe that there will be a positive relationship between SO and HR. We also noted the changes in baseball rules over the eras of baseball, and these rules may have affected the relationship between HR and SO between eras.

# 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 / Model

Descriptive Statistics:

The data was positively skewed, with 62.7% of players having hit 0 HRs and 21.8% of players having 0 SOs. Descriptive statistics show the 75th percentile for home runs hit was 5 HR’s (75% of players hit fewer than 5 HR’s in their careers), and 75th percentile for strikeouts was 105 SO’s (75% of players had fewer than 105 SO’s in their careers), compared to baseball’s leaders of 762 HR’s and 2597 SO’s.

Data Visualization:

The first step of our analysis was visualizing the raw data across all years (1871-2020) of HR versus SO for unique players to see how many HR’s and SO’s each player had accumulated throughout the history of baseball. The HR vs. SO scatterplot shows that many players are clustered to the bottom left of the plot (low HR, low SO) than players with the most HRs and SOs. The scatterplot does seem to show a positive linear relationship between HR and SO.

OLS Models (Non-Transformed Data):

The next step was to build an OLS linear regression model. Using all years of data, the OLS model of the overall relationship between HR and SO showed there was a strong relationship between HR and SO, with a model R-squared of 0.781. The p-value for the SO variable was 0.000 indicating a high degree of significance between HR and SO. According to the OLS model, for every 0.19 strikeouts, 1 HR will be hit. The errors were homoscedastic, and there was no autocorrelation. However, the distribution of the residuals was not normally distributed (Prob(Omnibus) 0.000, Skew 4.491, Kurtosis 73.384).

We then tested if baseball era would be significant in the model. When era was included in the model, the p-value for the era variable was 0.344 indicating that era did not significantly affect the relationship between HR and SO. Therefore, we did not include era in the overall OLS model.

While the OLS model with SO as the independent variable was significant with a good R-squared, the problem was that the distribution of the residuals was not normal. To try to address this problem, we assessed for outliers. Another OLS model was built, except this time outliers were removed. 272 outliers were removed but a similar relationship was found. The model without outliers had a R-squared of 0.827. The p-value for the SO variable was 0.000 indicating a high degree of significance between HR and SO. For every 0.17 strikeouts, 1 HR will be hit. The distribution of the residuals was still not normally distributed (Prob(Omnibus) 0.000, Skew 2.530, Kurtosis 41.464), but is closer to being more normally distributed than with the model with the outliers included.

While the OLS model without outliers was significant, with residuals more normally distributed, it may not be appropriate to completely remove the outliers. The outliers are players who could be considered “superstars” (the ones who hit many HR’s). Therefore, simply removing the outliers may not be a good idea. Therefore, we attempted to transform the data. We attempted many transformations (including Box Cox transformation), and the square root of both HR and SO produced the best results as the majority of the data is skewed to the right.

OLS Models (Transformed Data):

We ran 2 models of the transformed data (with and without outliers). The model with outliers had a R-squared of 0.819. The p-value for the SO\_transformed variable was 0.000 indicating a high degree of significance between HR and SO. Back transforming the model, for every 0.16 strikeouts, 1 HR will be hit. This value makes sense to be lower than the 0.17 coefficient in the OLS with outliers removed or 0.19 coefficient in the overall OLS, since the square root transformation attempts to compress the higher values, so lower values become more spread out. The distribution of the residuals was still not normally distributed (Prob(Omnibus) 0.000, Skew 0.581, Kurtosis 10.261) but much closer to normally distributed than the untransformed model without outliers.

The model of transformed data without outliers had a R-squared of 0.819. The p-value for the SO\_transformed variable was 0.000 indicating a high degree of significance between HR and SO. Back transforming the model, for every 0.14 strikeouts, 1 HR will be hit. The distribution of the residuals was still not normally distributed (Prob(Omnibus) 0.000, Skew 0.160, Kurtosis 7.676) but even closer to being normally distributed than all the other previous models.

In summary after comparing the various models for all years, the best model is the model using square root transformation of both HR and SO. It produced an excellent R-squared value, a statistically significant coefficient of the SO variable, and much more normally distributed residuals. It is probably best to include the outliers as we do want to account for the extremes of player talent (in both hitting home runs and also for striking out).

OLS Models by Era:

Lastly, we built OLS models for each of the 8 individual eras, with and without outliers, with and without square-root transformed data. While each era had a statistically significant p-value for the SO variable in the model, as the eras progressed, we see a stronger predictive relationship between SO’s and HR’s as evidenced by increasing R-squared values for all models.

|  |  |  |
| --- | --- | --- |
| **Era** | **Non-transformed R-Squared**  **(with outliers)** | **Non transformed R-Squared**  **(without outliers)** |
| Era 1 | 0.565 | 0.589 |
| Era 2 | 0.563 | 0.602 |
| Era 3 | 0.623 | 0.632 |
| Era 4 | 0.706 | 0.746 |
| Era 5 | 0.768 | 0.812 |
| Era 6 | 0.810 | 0.845 |
| Era 7 | 0.817 | 0.875 |
| Era 8 | 0.860 | 0.901 |

|  |  |  |
| --- | --- | --- |
| **Era** | **Transformed R-Squared**  **(with outliers)** | **Transformed R-Squared**  **(without outliers)** |
| Era 1 | 0.698 | 0.698 |
| Era 2 | 0.659 | 0.649 |
| Era 3 | 0.698 | 0.686 |
| Era 4 | 0.756 | 0.750 |
| Era 5 | 0.791 | 0.789 |
| Era 6 | 0.844 | 0.848 |
| Era 7 | 0.870 | 0.877 |
| Era 8 | 0.890 | 0.893 |

Non-OLS Models:

Although the square-root transformed data produced a reasonable OLS linear regression model, the non-transformed scatterplot did appear as though the data was linear and therefore might not be best to be transformed.

The main problem we encountered with the non-transformed model was the violation of the assumption of normality of residuals. This was likely because of the distribution of the data (predominance of low values of both HR and SO, right skewed). This raises the concept that perhaps an OLS linear regression model isn’t the best type of model to perform (even with removal of “outliers” which as previously mentioned may not be appropriate to remove as we do want to capture data from exceptional players). There are other types of regression models such Poisson, negative binomial, zero-inflated regression models that are better suited for this type of count data. These other regression models are beyond the scope of this assignment and this course, but are ones that could be considered in the future more advanced courses and analysis (we include some models in our Python notebook code, but without any interpretation in this report).

# Conclusions

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