# Data Science 2: Statistics for Data Science

## Summary Paper of the Analysis & Predictive Modelling of Strikeouts vs Home Runs in Major League Baseball Players 1871 - 2000

### Submitted by Group 8

Andrew Lagree

Rebecca Manning

Charlie Wang

Nathan Ye

Johanna Yu

### Supervised By:

Dr. Sergiy Nokhrin

Dr. Amanda Easson

Dr. Prakriti Sodhi

December 2021

#### Data Source

Statistical analysis has been a part of baseball since the inception of Major League Baseball (MLB). A century before it reached public consciousness, what would later be coined “sabermetrics” were being used in clubhouses to make key decisions about which players to draft, how to best utilize their skill set, and when to let them go. [[1]](#footnote-0)

[Lahman’s Baseball Database](http://www.seanlahman.com/baseball-archive/statistics/), which consists of pitching, hitting, and fielding data for Major League Baseball (MLB) from 1871 to 2020, is among the most widely used, accurate sources of MLB statistics on the internet. This database includes data from the two current MLB 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).

From [this database](https://github.com/chadwickbureau/baseballdatabank/archive/master.zip) we selected “Batting.csv” for our analysis because (1) batting statistics are of particular interest in MLB, as these metrics are fundamental to whether a team wins or loses games, (2) it is a rich data set, containing 108,789 lines of data, with batting statistics of 19,898 players across 149 years, and (3) the popularity of sabermetrics will allow for further research and comparison to reference models against which to evaluate our own.

#### Analysis & Model

“Batting.csv” includes player batting data across 22 variables. For our analysis we selected HRs and SOs as our variables of interest. Our hypothesis is that there is a significant, positive linear relationship between the number of home runs (HR) and number of strikeouts (SO) that a MLB player has over a given period of time. This hypothesis was based on the assumption that that players who hit more HRs swing the bat more aggressively and more often, which will also result in more strikeouts than more conservative batters. We will test this hypothesis over the entire history of baseball. To test our hypothesis, we chose to first do an exploratory analysis and visualization of the data, then to perform Ordinary Least Squares (OLS) regression to find an appropriate model for relationship.

To organize the data for analysis we performed 2 key steps: (1) merge results of the same player over multiple years into 1 line of data so that each line of data has career SO and HR data for each player and (2) oragnized the data by era of baseball to identify if trends and models created during our analysis were relevant to each era. The history of baseball can be broken into 8 specific eras, each with key differences in organizational structures and how the game was played[[2]](#footnote-1).

To visualize the relationship between SO and HR we generated a scatter plot of the 2 variables across the history of baseball. Though a positive linear relationship appeared to exist, we found that the data was heavily clustered around HR = 0. The data set proved to be positively skewed, where the 75th percentile of players had HR = 5 and SO = 105 compared to HR=762 and SO=2597 for baseball’s HR historical career HR leader.

Further exploratory analysis lead to observance of mean HR and SO by player over each of the 8 eras. We found that over time players SO more but have more success hitting HOs, and that their success in hitting HOs outpaces their increased SOs such that SO:HR decreases over the history of baseball.

Next, several OLS regressions were conducted to learn more about the relationship between HR and SO, and to find the optimal model. Model 1 used the original data and was used to explore if an OLS model would be suitable. It showed a strong relationship between HR and SO with R2 = 0.781 and predicted a SO:HR ratio of 5.2, such that 5.2 SOs would occur for 1 HR to be achieved. The errors were homoscedastic, there was 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. However, Kurtosis of 73.4 indicated that extreme outliers may be impact the normalcy of the distribution. Model 2, which involved removing 272 outliers and then performing the OLS regression, was conducted to improve on Model 1. Model 2 predicted an SO:HR ratio of 5.9, R2 = 0.827 indicates a better fit than Model 1 and Kurtosis of 41.5 showed that removal of the outliers from players with very high HR totals brought the data closer to being normally distributed than Model 2 but not yet fully satisfying this criteria.

A final OLS analysis was performed by transforming SO and HR using the square root function. This revealed a stronger correlation than the previous models, R2= 0.819. With Kurtosis = 10.3 this is the most suitable of the 3 models with respect to normalcy; further efforts to transform the data and better meet the normalcy condition were unsuccessful. After back transformation, the equation for the relationship of HR to SO with Model 3 was HR=0.1572 \* SO, such that the number of home runs could be predicted by the number of SOs at a rate of 6.36 SO to 1 HR.

Finally, we compared OLS regression for each of the 8 eras of baseball. As the eras progressed, we see a stronger relationship between SOs and HRs starting with R2 = 0.565 in era 1 and culminating with R2 = 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. This indicates that the effectiveness of the model varies from era to era and that a unique model may be required for each era.

#### Conclusions

We reject the null hypothesis that no relationship exists between number of HRs and SOs and conclude that that players who hit more HRs will have more SOs such that a positive, linear relationship exists between the two variables. Using a square root transformation of both variables and OLS regression, the best fit model found for the entire history of baseball HR can predict SO using the equation HR=0.1572 \* SO.

Further, we found that the fit of our OLS model improved over the 8 eras, such that while only 56.5% of the data from Era 1 fit the model, this increased era after era to 86% in Era 8. In other words, our model is a better fit for later eras of baseball and a poorer fit for the early years, indicating that though there is a clear, positive linear relationship between HRs and SOs in all 8 eras, that the differences in the game from era to era are significant enough to impact the SO:HR ratio such that a multiple predictive models by era as opposed to 1 model for the entire history of baseball would be a more accurate approach.

1. A guide to sabermetric research. Society for American Baseball Research. (n.d.). Retrieved December 5, 2021, from https://sabr.org/sabermetrics. [↑](#footnote-ref-0)
2. Woltring, T M, Rost, J K, Jubenville, C B. October 25, 2018. Examining Perceptions of Baseball’s Eras - Statistical Comparison. Retrieved from https://thesportjournal.org/article/examining-perceptions-of-baseballs-eras/ [↑](#footnote-ref-1)