HITS & HOMERS: A STATISTICAL LOOK AT MLB's BEST HITTERS

STOCKTON NELSON
and
ISAAC MILLER
and
COOPER RIGGS

April 14, 2025

Abstract

In Major League Baseball, teams are always seeking ways to gain a competitive edge. By evaluating various variables in a player's profile, it's possible to predict how he is likely to contribute on the field. We wanted to answer two key questions. First, how are offensive performance metrics—such as slugging percentage, on-base percentage, plate appearances, and strikeout rate—related to both the number of home runs and a player's batting average over a season? Our other question was, how do the factors associated with hitting more home runs differ from those associated with achieving a higher batting average, and what does this reveal about the distinct skills involved in power hitting versus consistent contact? To answer these, we created two models using a large dataset of hitting statistics from Baseball-Reference. We found that many of the predictors in the dataset were highly correlated, which initially hindered model accuracy. However, after accounting for multicollinearity, we developed models that incorporate numerous quantitative and a few categorical variables to effectively predict a player's batting average and home run totals.

1 Introduction

In baseball, offensive success is often summarized by just a few key statistics—most notably, a player's batting average and number of home runs. While both metrics are used to evaluate hitting performance, they reflect distinct aspects of a player's offensive skill: batting average captures consistency in making contact and reaching base, while home runs represent raw power and the ability to change the game with a single swing. Understanding what influences each of these outcomes is valuable not only for fans and analysts but also for coaches, scouts, and front offices aiming to evaluate and develop talent more effectively. In this study, we explore how a range of offensive performance measures—such as slugging percentage, on-base percentage, plate appearances, and strikeout rate—relate to both batting average and home run totals across MLB players. Specifically, we investigate: (1) how offensive performance metrics are related to both the number of home runs and a player's batting average over a season; and (2) how the factors associated with hitting more home runs differ from those associated with achieving a higher batting average, and what this reveals about the distinct skills involved in power hitting versus consistent contact.

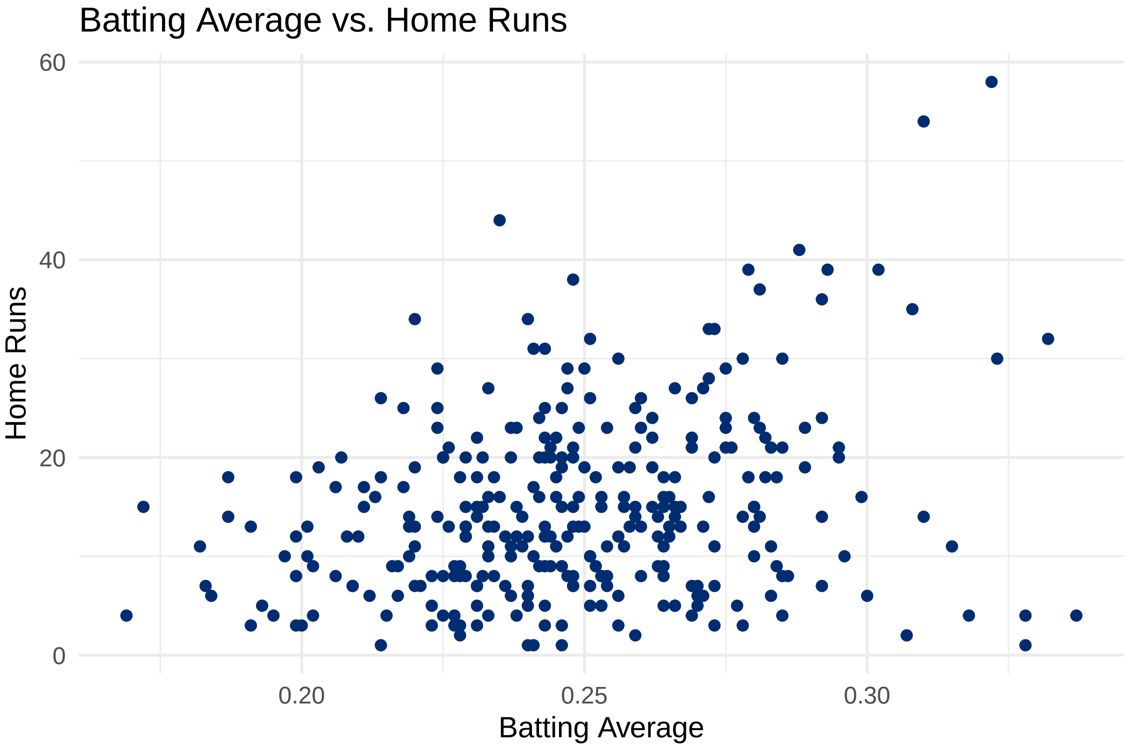
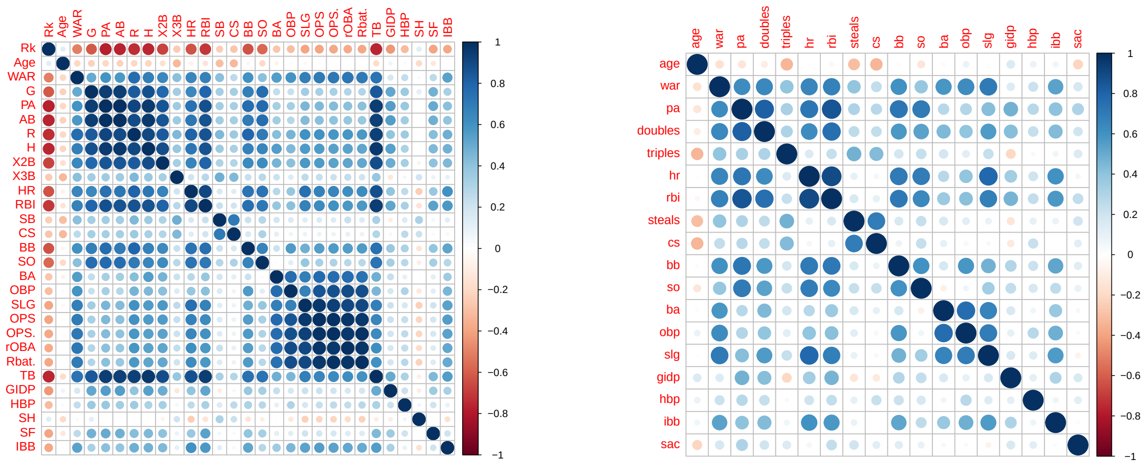


Figure 1: Scatterplot of Home Runs by Batting Average.

1.1 Data Description

This dataset was obtained from Baseball Reference and includes season-long batting statistics for the top 300 Major League Baseball (MLB) players in the 2024 regular season, ranked by batting average. Each row in the dataset represents a single player's full-season batting performance, making the observational unit a player-season. The variables capture a range of offensive statistics, including measures of hitting efficiency (e.g., slugging percentage and on-base percentage), power hitting (e.g., home runs, doubles, and triples), and baserunning (e.g., stolen bases and total bases). Most variables are either raw counts (e.g., home runs, doubles, stolen bases) or percentages (e.g., batting average, slugging percentage). These statistics provide a detailed view of player performance and are used to investigate which aspects of offensive play are most closely associated with outcomes like batting average and home run totals.

To better understand the nature of the data, it's helpful to know that most hitters in this group have a batting average between .200 and .270, with standout performers exceeding .300. In terms of home runs, most players hit between 8 and 20 in the season, while top power hitters reach 30 or more. Additionally, many of the offensive statistics in baseball are variations or combinations of one another, resulting in high correlations among predictors. This multicollinearity raised concerns during modeling and motivated the use of regularized regression methods such as LASSO.



(a) All variables Correlation Matrix

(b) Updated Correlation Matrix

Figure 2: Side-by-side correlation matrices showing (a) the correlation matrix with all the variables in the dataset and (b) an updated correlation matrix after getting rid of correlated variables.

2 Model Selection and Validation

For the study, our goal was to select the best models possible to predict batting average and home runs. To do this, we first selected models ourselves that we thought would work to predict batting average and home runs. We then ran AIC and BIC best subsets, and LASSO each twice on our large dataset of baseball statistics, once to create predictor models for batting average and once to create predictor models for home runs, thus creating four candidate models each for predicting batting average and for predicting home runs. Finally, we ran cross validation twice, once on the batting average models to select the best of the four options for batting average and once on the home run models to select the best for home runs.

Here we will show first the model we originally tested for batting average, then the one we selected for batting average after running cross-validation. Then we will show the same for home runs.

This is our test model for batting average. We chose this because we felt that a player that walks a lot and gets on base a lot also would likely have a high batting average. Similarly, getting lots of RBI usually indicates getting a lot of hits, which would also contribute to having a higher batting average:

$$\text{BattingAverage}_i = \beta_0 + \beta_1 \times \text{Walks}_i + \beta_2 \times \text{RBI}_i + \beta_3 \times \text{OBP}_i + \epsilon_i, \tag{1}$$

where

$$\epsilon_i \sim N(0, \sigma^2).$$

After running cross-validation, the model we ultimately selected to predict batting average was the LASSO-produced model, which is as follows:

$$\begin{aligned} \text{BattingAverage}_i = \beta_0 + \beta_1 \times \text{WAR}_i + \beta_2 \times \text{PlateAppearances}_i + \beta_3 \times \text{Doubles}_i + \beta_4 \times \text{Walks}_i \\ + \beta_5 \times \text{Strikeouts}_i + \beta_6 \times \text{OBP}_i + \beta_7 \times \text{SLG}_i + \beta_8 \times \text{GroundedIntoDoublePlay}_i + \beta_9 \times \text{HitByPitch}_i \\ + \beta_{10} \times \text{SacFlies}_i + \beta_{11} \times \text{I}(\text{Divison}_i = \text{ALWest}) + \beta_{12} \times \text{I}(\text{Division}_i = \text{NLEast}) + \epsilon_i, \end{aligned} \tag{2}$$

where

$$\epsilon_i \sim N(0, \sigma^2).$$

From the cross validation we did for Batting Average models, the LASSO-picked model came out to be the one with the lowest bias, and had the same amount of predictors as the AIC model; thus, we chose to use this model for batting average.

This is our test model for predicting home runs. We chose these predictors because we thought that having a high batting average could indicate that a hitter would also be good at hitting home runs; SLG and RBI usually indicate a player getting lots of extra-base hits; strikeouts indicate a player taking more risks in at-bats; and Walks indicate a hitter being able to see well where the ball is:

$$\begin{aligned} \text{HomeRuns}_i = \beta_0 + \beta_1 \times \text{BattingAverage}_i + \beta_2 \times \text{SLG}_i + \beta_3 \times \text{RBI}_i \\ + \beta_3 \times \text{Strikeouts}_i + \beta_3 \times \text{Walks}_i + \epsilon_i, \end{aligned} \tag{3}$$

where

$$\epsilon_i \sim N(0, \sigma^2).$$

The Model we chose for predicting home runs, after cross-validation:

$$\begin{aligned} \text{HomeRuns}_i = \beta_0 + \beta_1 \times \text{WAR}_i + \beta_3 \times \text{Doubles}_i + \beta_4 \times \text{Triples}_i \\ + \beta_5 \times \text{RBI}_i + \beta_6 \times \text{Walks}_i + \beta_7 \times \text{Strikeouts}_i \\ + \beta_8 \times \text{OBP}_i + \beta_9 \times \text{SLG}_i + \beta_{10} \times \text{HitByPitch}_i \\ + \beta_{11} \times \text{SacrificeHits}_i + \beta_{12} \times \text{IntentionalWalks}_i \\ + \beta_{13} \times \text{I}(\text{Division}_i = \text{NLCentral}) + \epsilon_i, \end{aligned} \tag{4}$$

where

$$\epsilon_i \sim N(0, \sigma^2).$$

Based on the cross validation for our Home Run models, the AIC model clearly had the lowest prediction error (PMSE) and the closest-to-zero bias; thus, we chose to use the AIC model to predict home runs.

We had originally selected models using predictors from the entire dataset; however, upon checking for assumptions, we found that several were not being met—including multicollinearity, normality, influential points, and equal variance. Because of this, we had to remove some variables from the dataset and re-do the process of selecting our models, which resulted in models that fit the assumptions. One way we also addressed this was by using LASSO regression, which improves model accuracy and interpretability by shrinking some coefficients to zero, effectively performing variable selection. Unlike stepwise or best subsets methods, LASSO can handle many predictors at once and is more stable in the presence of multicollinearity, making it more useful for building simpler, more generalizable models.

3 Analyses, Results, and Interpretation

Table 1: Regression Output for Batting Average Model

	Estimate	t value	p-value	Lower 95%	Upper 95%
(Intercept)	-0.00866	-0.96595	0.33485	-0.02631	0.00899
WAR	0.00274	5.12147	0.00000	0.00169	0.00380
Doubles	0.00101	8.84421	0.00000	0.00078	0.00123
BB	-0.00107	-18.51070	0.00000	-0.00118	-0.00095
SO	0.00002	0.69392	0.48827	-0.00003	0.00007
OBP%	0.00808	23.28547	0.00000	0.00740	0.00876
SLG%	0.00030	1.82021	0.06973	-0.00002	0.00062
GIDP	0.00071	4.83644	0.00000	0.00042	0.00101
HBP	-0.00139	-8.49005	0.00000	-0.00171	-0.00107
SAC	0.00088	3.87427	0.00013	0.00043	0.00132
AL West	0.00004	0.02260	0.98198	-0.00331	0.00339
NL East	0.00134	0.82068	0.41248	-0.00187	0.00456

The results of the batting average model (Table 1) show us some interesting things about contributions to a player's batting average in the most recent season. The most influential predictor was on-base percentage. With a very low p-value we can assume that on-base percentage is significant in predicting batting average. In fact, we are 95% confident, holding all other variables constant, that for every .010 increase in a player's on-base percentage his batting average will increase, on average, between .007 and .009. Due to the narrowness of this interval we can be very confident in on-base percentage as a predictor of batting average. This makes sense because when a player gets a hit it contributes positively to both their batting average and their on-base percentage. Interestingly, the results show that walks are negatively associated with batting average. We are 95% confident that, on average, for every walk a player has their batting average will decrease between .001 and .0009, holding all the other variables constant. While this isn't a large change in batting average it is still a very narrow interval, which shows that walks are a significant contributor in decreasing a player's batting average. Other results we found are that slugging percentage, while in the prediction model, gives us 95% confidence, while holding all else constant, that for every .010 increase in slugging percentage a player's batting average changes between -0.00002 and 0.0006, on average. This means that we cannot claim that slugging percentage has a real effect on batting average. This is interesting because with 95% confidence we can say that, holding all else constant, for every double a player hits their batting average will increase, on average, by 0.0007 and 0.001, showing that doubles has a significant effect on batting average while slugging percentage, which doubles contributes to, does not have a significant effect. From these results it can be safe to say that players that make consistent contact, rather than hitting for power are more likely to increase their batting average.

Table 2: Regression Output for Home Run Model

	Estimate	t value	p-value	Lower 95%	Upper 95%
(Intercept)	1.654	1.030	0.304	-1.507	4.814
WAR	0.455	4.665	0.000	0.263	0.647
Doubles	-0.179	-8.783	0.000	-0.219	-0.139
Triples	-0.444	-7.936	0.000	-0.554	-0.334
RBI	0.147	13.298	0.000	0.125	0.168
BB	0.167	14.238	0.000	0.144	0.191
SO	0.027	5.897	0.000	0.018	0.036
OBP%	-1.161	-17.222	0.000	-1.294	-1.029
SLG%	0.884	24.524	0.000	0.813	0.955
HBP	0.154	5.204	0.000	0.096	0.212
IBB	0.182	3.221	0.001	0.071	0.293
SAC	-0.137	-3.255	0.001	-0.220	-0.054
NL Central	-0.465	-1.632	0.104	-1.026	0.096

The results of our Home run model (Table 2) show us what contributes to a player's home run total in the most recent season. We can see from the results that slugging percentage is the most important factor in a player's home run total. We are 95% confident, while holding all else constant, that an increase of .010 in a player's slugging percentage, will increase their home run total between .813 and .955, on average. This shows that slugging percentage contributes greatly to a player's home run total. Another interesting result is that all types of walks (BB, HBP, and IBB) are all positively significant contributors to home run total. For example, with 95% confidence, holding all the other variables constant, we can say that for every walk a player's home run total, on average, will increase between 0.144 and 0.191. Over the course of the season walks can start to add up, so it's interesting to see that a player who has 80 walks in a season would, on average, have a home run total 11.5 to 15 home runs higher. Not only do walks contribute positively to home run total, but we also found that we are 95% confident, holding all else constant, for every strikeout a player's home run total will increase, on average, between 0.018 and 0.036. While it is a smaller number than walks, it is interesting to see that strikeouts have a significant positive effect on home run total. Finally, we found that doubles, triples, and on-base percentage, while assumed to have a positive effect on home run total actually have a negatively significant effect on home run total, with each of these variables having very small p-values we have sufficient evidence to claim this. In fact, with 95% confidence we claim, while holding all else constant, that for every triple hit a player's home run total will decrease, on average, between 0.334 and 0.554 home runs.

As we looked at the results of our home run model, we noticed that there were a couple of somewhat influential points in the model. These points were two players who led the MLB in home runs last season, Aaron Judge and Shohei Ohtani. We were curious if our model performance would improve if these two players were removed from the model. Thus, we removed them and ran a new model without Judge and Ohtani. With the results of our first model (Table 2) and the results of our new model without the influential players we compared predictive metrics such as adjusted R^2, RMSE, AIC, and BIC. What we found was that the new model produced lower RMSE, AIC, and BIC. For example, the AIC for the first model was 1273.6 and for the new model it was 1249.8, with the difference being 23.8 showing that there is a significant difference in the two models, suggesting that the new model is better at predicting home run totals than the first model. We also found that the new model had an RMSE of 1.76 and the first model had an RMSE of 1.80, showing that the new model, on average, predicted better than the first model. While the new model didn't predict much better than the first model, our predictive metrics still showed that the new model performed better. This allows us to conclude that Shohei Ohtani and Aaron Judge overachieved in their home run totals given their other offensive performance metrics, justifying both players winning MVP in their respective leagues last season.

4 Conclusions

Overall, through our analysis we made several interesting findings about what statistics most predict batting average and home runs for a given baseball player in a season. Originally, our questions were about finding these variables, and also whether the predictors for batting average and home runs are different, or if many of their predictors overlap. We found that for batting average, OBP was the most influential predictor, with walks as the second most influential, negatively influencing batting average. This was interesting because walks contribute to OBP as well. Notably, doubles also contributed strongly to batting average. For home runs, RBI, Walks, and SLG were the strongest that contirbuted positively, while OBP was the most influential predictor that cotributed negatively. One variable that positively predicted both was WAR, however, it seemed that most of the variables that positively influenced batting average, negatively influenced home runs, and vice versa. From this we can conclude that overall the predictors for batting average and home runs are largely different, showing that overall the required skillsets in players are quite different for high batting average and high home runs, respectively. However there are some weaknesses with our models that could be problematic for our results. For example, grounding into double plays was a significant positive predictor for batting average in our analysis. That said, clearly grounding into double plays should not be recommended for any player to improve their batting average. In a similar fashion, we found strikeouts were a significant positive predictor for home runs. While many of the predictors we found are valuable to know for predicting batting average and home runs, it is still important to look at the results logically.

APPENDIX

Prompts:

- I am reporting on a regression and I am looking for feedback on my report. I have a few things I want to look over, so I will go over each of them one by one. To start, do the ideas connect throughout the paper?
- My report calls to write in the context of writing to a boss who has taken STAT 121 and STAT 330 and is familiar with some statistical concepts. For reference, STAT 121 is an intro to statistics class and STAT 330 is an introduction to linear regression. Does the report with the changes we've made fit into that context?
- Are the research questions proposed in the report answered?

Recommendations:

- Page 4 (Batting Average model explanation): The switch between the original model and the LASSO model could be more explicitly contrasted. You could say something like: ”While the initial model relied on intuition (e.g., Walks and RBIs), LASSO offered a broader, data-driven selection that improved predictive performance.”
  - **How We Incorporated It:** Revised the beginning of this paragraph to flow better.
- Conclusion: This section is solid, but one sentence that begins with “However there are some weaknesses…” runs on a bit. Consider splitting or tightening it to make the ideas punchier.
  - **How We Incorporated It:** Revised this section to be split up better.
- You could soften one or two heavier stats terms (like “PMSE”) by quickly saying what it means: “...the AIC model clearly had the lowest PMSE (prediction error)...”
  - **How We Incorporated It:** We revised it so our boss can understand what we are saying.
- You might also add brief intuitive reasoning behind LASSO in the model selection section (e.g., “We chose LASSO because it works well when predictors are highly correlated, helping simplify the model without losing accuracy.”)
  - **Why We Didn't Incorporate It:** We already had a section at the beginning of page 4 which covers this.
- Both research questions are clearly answered.
  - **Why We Didn't Incorporate It:** Says that we have answered both questions.