Determining the Best Means to Predict Single-Season Home Run Totals of Major League Baseball (MLB) Hitters

Introduction

The research question of which metrics are best at predicting home runs using linear regression is critical and novel in the field of baseball analytics. The rationale behind this research is to gain knowledge about which factors are crucial in determining a player's home run performance. This is important because it will determine the ability to generate offensive production in the sport. Put simply, home runs are an efficient way to generate runs, and over the past several decades, an increased emphasis has been placed on building offenses to produce a greater percentage of their runs via the home run.

According to Samford University based on the last ten seasons the number of home runs is positively correlated with the number of games won, with a $r = 0.37$ (Williams, 2019). Similar analyses have produced an r-value of 0.42 (Prusaczyk, 2016). While the exact relation changes year to year, the percentage of a team's runs being scored by home run, as well as the total number of home runs hit, are perennially important benchmarks when comparing to overall team success.

Not only are generating home runs important from the team perspective, they are also crucial to making the sport more exciting – with increased rates of home runs in recent years leading to allegations that MLB itself has contributed to the surge by changing the physics of the baseball itself (Albert et al., 2018).

Breaking down from the total number of home runs and wins, the individual components of each player are important underpinnings of a team's success. Internally, being able to accurately predict the offensive contribution of each player allows a team flexibility in how they can structure their roster over the course of a season.

Our research is important because it will enhance the existing understanding of the complex relationships which intersect to determine a player's home run performance. In turn, this information can be used by coaches and general managers to choose players, trade players, and develop players. This will lead to more cost effective allocation of available resources to be used on players which will lead to increased efficiencies in the league along with competition. Some questions our team had going into the research were:

1. **Do right handed batters hit more home runs than left handed batters?**
2. **Are advanced statistics (wRC+, Exit Velocity, Launch Angle) a better predictor of home runs than traditional statistics (Batting Average, Strikeout Percentage)?**
3. **Do outfield players hit more home runs than infield players?**

Data Summary

Our data was collected by BaseballSavant (an MLB product), Baseball-Reference, and Fangraphs during the 2022 Major League Baseball (MLB) season. Most advanced statistics, such as Exit Velocity and Launch Angle were collected by use of StatCast, a measurement system which has been in place in each MLB ballpark since 2015, and utilizes a system of cameras and sensors to detect movement and spin on the baseball. Specifically, our observations are of the top 100 home run hitters during the 2022 MLB season and their accompanying explanatory variables. Our quantitative variables are wRC+ (Weighted Runs Collected Plus), Exit Velocity, Launch Angle, Strikeout Percentage and Batting Average and were collected from each individual player's season averages from the 2022 MLB season. Some of our qualitative explanatory variables change season upon season. These variables are which league the player plays in and the positions of the players. For these qualitative variables, we used their 2022

statistics as our variables. Specifically, for the qualitative variable of League, we used the League in which the player played the most games in that season to account for mid-season trades between leagues. Likewise, for position, we used the position in which that player played the most games in during the 2022 season. Our final qualitative variable is Dominant Hand, the side of the plate from which a batter hit, which was obtained via Baseball-Reference.

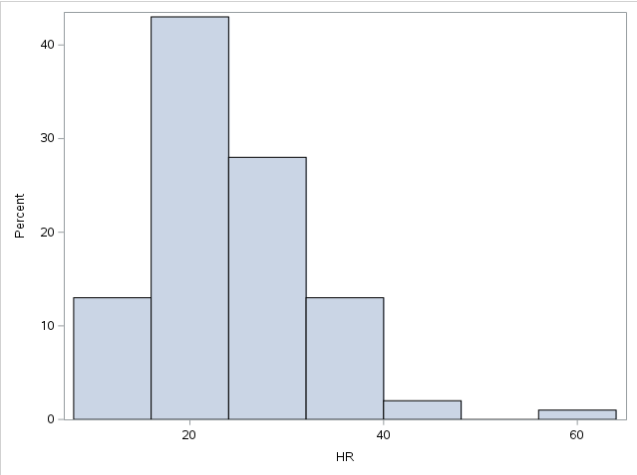
One concern regarding the data collected is the fact that each player in our sample did not play the same amount of games in the 2022 MLB season. The amount of games played in our sample may be different for many reasons including but not limited to injury, personal leave, and relocation after being traded. We believe that this difference in games played in our sample may lead to discrepancies in the amount of home runs hit as well as a subset of our quantitative explanatory variables. Further, there may be players who otherwise would have been listed in the top 100 home run hitters who were hampered by injury, and thus their metrics were excluded for the purposes of this analysis.

Along with this, we are concerned about how accurate our model will be for the population of all MLB players if our regression model is applied to them. Since our sample only includes the top 100 home run hitters during the 2022 MLB season, we foresee that our model may be slightly skewed when applied to the many MLB players who hit between 0 and 10 home runs over the course of a 162 game schedule. Further, the distribution of home runs hit by top players is roughly normal. The distribution of home runs hit by *all* players (including those with only a few games played) is heavily left skewed.

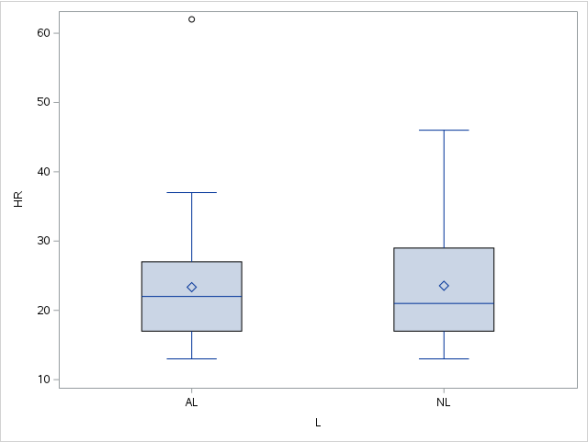
We can confidently trust our sources because baseball statistics are all a standardized measurement, and all statistics are based purely on player performance. We also can safely assume that there are no external factors influencing results (i.e. cheating) in the form of performance enhancing drugs or outside knowledge in the form of sign stealing, for instance. All statistics were collected from sources that are directly tied to MLB (BaseballSavant, Baseball-Reference), or sources that are MLB-adjacent, but trusted to the extent that they are used in contract negotiations (Fangraphs).

Exploratory Data Analysis

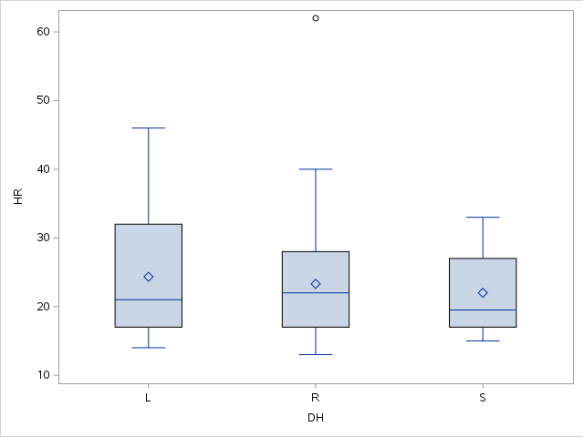
Histogram of Home Runs of the Top 100 Home Run Hitters in the 2022 MLB Season



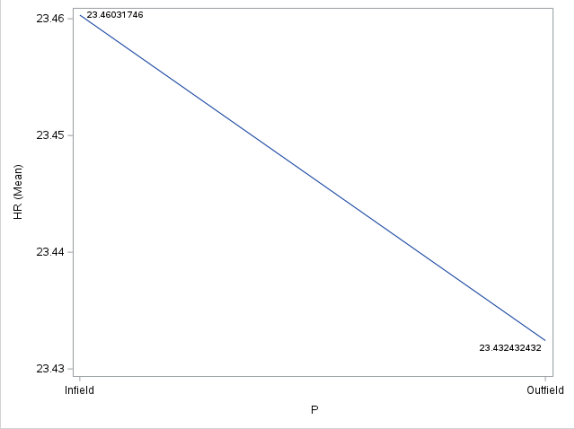
Mean of Home Runs Across Leagues



Mean of Home Runs Across Handedness



Mean of Home Runs Across Position



Correlations of Advanced Statistics with the Response Variable (Home Runs)

The CORR Procedure

1 With Variables:	HR
3 Variables:	EV wRC LA

Pearson Correlation Coefficients, N = 100			
HR	wRC 0.60443	EV 0.55030	LA 0.32848

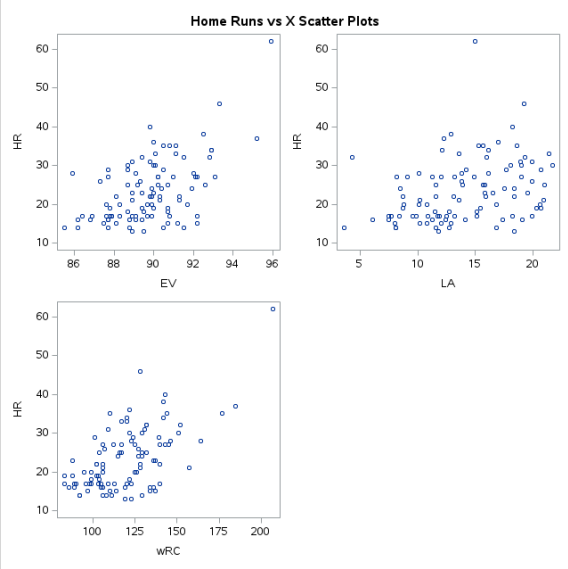
Correlations of Traditional Statistics with the Response Variable (Home Runs)

The CORR Procedure

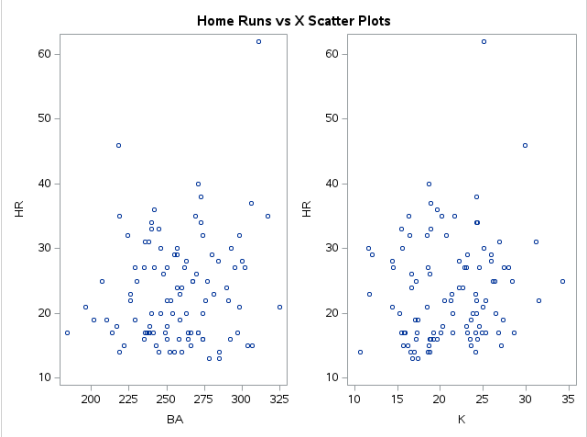
1 With Variables:	HR
2 Variables:	K BA

Pearson Correlation Coefficients, N = 100		
HR	BA 0.14847	K 0.14318

Scatter Plots of Advanced Statistics with Home Runs



Scatter Plots of Traditional Statistics with Home Runs



Correlations of Explanatory Variables with One Another

Pearson Correlation Coefficients, N = 100 Prob > |r| under H0: Rho=0					
EV	EV 1.00000	wRC 0.43170 <.0001	K 0.30886 0.0018	BA 0.14610 0.1469	LA -0.08432 0.4042
LA	LA 1.00000	BA -0.38601 <.0001	K 0.13497 0.1806	EV -0.08432 0.4042	wRC -0.03994 0.6932
K	K 1.00000	BA -0.42770 <.0001	EV 0.30886 0.0018	wRC -0.29487 0.0029	LA 0.13497 0.1806
BA	BA 1.00000	wRC 0.72847 <.0001	K -0.42770 <.0001	LA -0.38601 <.0001	EV 0.14610 0.1469
wRC	wRC 1.00000	BA 0.72847 <.0001	EV 0.43170 <.0001	K -0.29487 0.0029	LA -0.03994 0.6932

Conclusion

After conducting Exploratory Data Analysis, we found that our response variable of Home Runs during the 2022 MLB season is suitable for regression as the spread of the variable is continuous (one outlier), unimodal and relatively symmetric. We also ran a collection of analyses to better understand how each of our explanatory variables were related to home runs and to each other. As shown above via box and whisker plots, there is little difference between the two leagues in the mean number of home runs hit, nor in the distribution of those home runs. Similarly, the mean and spread of home runs across handedness (Right, Left, or Switch) demonstrated no distinct differences. These conclusions may be reached by observing the large degree of overlap between the means and 25 to 75 percentiles.

In addition to the lack of difference by league and handedness, the mean number of home runs hit by position (reduced to simple infield/outfield distinctions) were almost exactly equivalent (23.46 for infielders, 23.43 for outfielders).

To address our second question, however, we compared the correlations between home runs and quantitative advanced statistics (wRC, EV, LA) to the correlations between home runs and quantitative traditional statistics (BA, K). As shown in the output above, the correlation between advanced statistics was significantly higher, headlined by wRC and EV with Pearson correlation coefficients above 0.55. Conversely, BA and K each produced Pearson scores of about 0.15, less than half of even the most loosely correlated advanced statistic, LA. These correlations are further supported visually by the scatter plots listed after the numeric interpretations of these relationships.

Observing the scatter plots of each individual explanatory variable with our response variable of home runs, we can see that LA, B and K seem to have no correlation with our response variable. On the other hand, it appears that the explanatory variables of wRC and EV have a moderately positive linear relationship with home runs. This graphical analysis goes hand in hand with our analysis of the correlation values between explanatory variables and the response variable.

Finally, to begin initial investigation into multicollinearity between our variables, we produced a matrix of each quantitative explanatory variable and every other quantitative explanatory variable. While there were not any relationships which were particularly strong (above a benchmark of 0.75, for instance), the correlation between wRC and BA came close at 0.73. Otherwise, wRC and EV produced a correlation of 0.43, while BA and K were inversely related, as evidenced by a -0.43 Pearson correlation coefficient. The rest of our variables exhibited even weaker relationships, though this does not entirely quell concerns surrounding multicollinearity.

These initial results provide strength to our question of whether advanced statistics are a better predictor of home runs than traditional ones, though they are not conclusive. To move forward, we will plan on collecting more variables in each of these categories for our current sample of players. Further, the relative weights (t test, for instance) of each of the predictive power of each of these variables is yet to be determined. In building our model, we will be able to further investigate whether traditional statistics are valuable pieces of information at all.

In terms of our qualitative variables, we saw little difference between home run means by handedness, league, and position. These results make sense, as there are not large disparities in each of these bins which have ingrained themselves in the baseball landscape. However, we will continue to explore these variables, as well as new ones (e.g. Free Agent) to supplement our quantitative analysis with categorical distinctions.

Appendix 1 - Data Dictionary

Variable	Referred to As	Description	Units
Home Runs	HR	The number of Home Runs hit (i.e. the number of balls hit over the outfield wall)	Number of Home Runs
Dominant Hand	DH	Which side of the plate the batter hits from	Right/Left/Switch
Position	P	Whether the player plays in the infield or outfield	Outfield/Infield
League	L	Which league in the MLB that the player played in (American League or National League)	AL/NL
Weighted Runs Created Plus	WRC	WRC+ takes the statistic Runs Created and adjusts that number to account for important external factors -- like ballpark or run-scoring environment. It's adjusted, so a WRC+ of 100 is league average and 150 would be 50 percent above league average	Unitless (Scaled so 100 is Average)
Exit Velocity	EV	The speed at which the ball is launched off the batters bat	Miles per Hour
Launch Angle	LA	The angle at which the ball is launched off of the batters bat	Angle
Strikeout Percentage	K	Ratio of player's strikeouts to official plate appearances	Percentage
Batting Average	BA	The number of hits per 1000 at-bats	Number

Appendix 2 - Data (First 20 Observations)

Name	HR	DH	P	L	wRC	EV	LA	K	BA
Aaron Judge	62	R	Outfield	AL	207	95.9	15.0	25.1	311
Kyle Schwarber	46	L	Outfield	NL	128	93.3	19.2	29.9	218
Pete Alonso	40	R	Infield	NL	143	89.8	18.2	18.7	271
Austin Riley	38	R	Infield	NL	142	92.5	12.9	24.2	273
Yordan Alvarez	37	L	Outfield	AL	185	95.2	12.3	18.9	306
Christian Walker	36	R	Infield	NL	122	90.0	17.0	19.6	242
Mookie Betts	35	R	Outfield	NL	144	90.5	18.6	16.3	269
Paul Goldschmidt	35	R	Infield	NL	177	90.8	15.7	21.7	317
Rowdy Tellez	35	L	Infield	NL	110	91.1	15.3	20.2	219
Matt Olson	34	L	Infield	NL	120	92.9	16.1	24.3	240
Shohei Ohtani	34	L	Outfield	AL	142	92.9	12.1	24.2	273
Anthony Santander	33	S	Outfield	AL	120	90.1	21.4	18.9	240
Corey Seager	33	L	Infield	AL	117	91.1	13.6	15.5	245
Anthony Rizzo	32	L	Infield	AL	132	89.4	19.3	18.4	224
Manny Machado	32	R	Infield	NL	152	91.5	16.0	20.7	298
Vladimir Guerrero Jr.	32	R	Infield	AL	132	92.8	4.3	16.4	274
Eugenio Suarez	31	R	Infield	AL	131	89.8	19.9	31.2	236
Willy Adames	31	R	Infield	NL	109	88.9	18.9	26.9	238
Kyle Tucker	30	L	Outfield	AL	129	90.0	19.0	15.6	257
Nolan Arenado	30	R	Infield	NL	151	88.7	21.7	11.6	293

Data Taken and Compiled from Baseball Savant, Baseball-Reference and FanGraphs

Works Cited

Albert, J., Bartroff, J., Blandford, R., Brooks, D., Derenski, J., Goldstein, L., Hosoi, A., Lorden, G., Nathan, A., & Smith, L. (2018). *Report of the Committee Studying Home Run Rates in Major League Baseball*, 1–84. https://doi.org/http://baseball.physics.illinois.edu/HRReport2018.pdf

FanGraphs. (n.d.). *RosterResource*. Fangraphs Baseball. Retrieved March 20, 2023, from https://www.fangraphs.com/

MLB Advanced Media. (n.d.). *Baseball savant: Trending MLB players, Statcast and Visualizations*. baseballsavant.com. Retrieved March 20, 2023, from https://baseballsavant.mlb.com/

Nath, A. M. (n.d.). *The Physics of Baseball* . The Home Run Surge. Retrieved March 20, 2023, from http://baseball.physics.illinois.edu/homeruns.html#:~:text=The%20increase%20in%20home%20runs,resulting%20in%20more%20home%20runs

Prusaczyk, J. (2016, September 9). *Do more home runs mean more wins?* Beyond the Box Score. Retrieved March 20, 2023, from https://www.beyondtheboxscore.com/2016/9/9/12842846/home-runs-wins-correlation-2016

Sports Reference. (n.d.). *MLB Stats, Scores, History, & Records*. Baseball. Retrieved March 20, 2023, from https://www.baseball-reference.com/

Williams, C. (2019, June 17). *How important are home runs in a power hitting period?* Samford University. Retrieved from https://www.samford.edu/sports-analytics/fans/2019/How-Important-Are-Home-Runs-in-a-Power-Hitting-Period