Heavy Hitters & Breadwinners: A Statistical Analysis on MLB Hitter Salaries

Max Pyshniak, Joel Lawrence

MacEwan University

STAT 378: Applied Regression Analysis, AS01

Instructor: Dr. Rui Hu

December 7, 2022

Abstract

Baseball is one of the most popular sports in the world, and along with it come some of the highest salaries. To investigate what really matters in terms of hitter salaries, multiple aspects must be considered. Through the usage of various model selection techniques, it was found that the selection made by the Akaike Information Criterion was preferable to that of the Bayesian Information Criterion or the LASSO method. This decision was made largely based with a focus on the adjusted coefficient of determination, which is a more concrete measure of performance than the AIC or BIC individually. A proposed multiple linear regression model was created to analyze the statistics of the 2021 active players in the MLB and highlighted which features to be important in determining a hitter's annual income.

Keywords: *Multiple Linear Regression, Statistics, Major League Baseball [MLB], Hitters, Salaries*

Heavy Hitters & Breadwinners: A Statistical Analysis on MLB Hitter Salaries

Introduction

Baseball is known as one of North America's favourite pastimes, and watching it played on the professional level in the MLB is no exception for sports fanatics. Many people could only dream that the activity they played casually with their friends at a young age would turn into one of the highest earning careers. The skill, game knowledge, and determination that it takes to become a professional baseball player is held by only a handful of talented individuals. With many factors that can determine a player's success, statistical modelling can reveal which variables are significant in predicting their salary, a strong motivation towards their success. If a player performs better on the field, the amount that they are getting paid may reflect that. Although there are varying techniques in baseball, this paper aims to produce an effective salary prediction model specifically for batting analytics**.** To better understand what is being analyzed, the rules and standards of baseball had to be researched. Each team takes turns playing on defense (involving pitching, catching and fielding) and on offense (involving batting and running) through nine innings. Batting is the driving force for a team scoring a run against their opponent. If a player excels at batting the ball pitched to them, the greater the run production for their team, proving the player is valuable. The hitter bats the ball towards the field and the farther they bat, the more bases they can pass, bringing them closer to scoring a run. The team on defense will attempt to stop the runner from passing each base by either catching the ball before it touches the ground, tagging them with the ball, or returning the ball to a base before they can reach it.

Many variables that were taken into consideration to predict a batter's salary. Widely accessible hitting statistics from the 2021 season for each player, released by the MLB include hits, at-bats (times a player came up to the plate to bat), batting average (number of hits divided by at-bats), bases on balls (walks), and games played. Runs are counted as the number of times a player reaches home plate, and "runs batted in" are any runs scored from a player batting in a single play. Along with the 2021 season, the listed statistics were also collected from the start of each player's career, to give insight into their performance and legacy as a whole. Given that batters also play on the field while they are on defence, it was crucial to examine their fielding performance including putouts (physically putting a player out), assists (assisted in putting a player out) and errors (making a mistake that benefits the offensive team). Variables outside of the standard statistics of baseball were harvested such as the hitter's height and weight. Batting and running take a lot of raw strength, therefore a person's size may influence how well they can play. Furthermore, categorical information such as team and league were taken as factors that may affect salary.

This report's goal is to obtain the factors that may predict an MLB hitter's salary. With the use of multiple linear regression models, the analysis of a player's performance and background can reveal which variables are statistically significant to the money they earn. Based on the MLB Stat Leaderboard from ESPN (*2022 MLB stat leaders*), the desired standard statistics in baseball appeared to be batting average, home runs, runs batted in, and hits. Therefore, a hypothesis was made to predict these specific variables to be the most valuable and have the greatest influence on a hitter's salary. In addition, the statistics from a player's entire career rather than a single season were predicted to be more significant as one's legacy would be expected to garner them more market value.

Methods

Data Collection

Data was amalgamated using tables provided by the Chadwick Baseball Bureau (Chadwickbureau). All rows for the given year (2021) were extracted from a table in the file Batting.csv, to get the player IDs of all active players in that season. Then, all prior years of those player IDs were added to the table as well. Joining these rows with “Fielding.csv” created a table with the complete player history for those who were active and showed evidence of hitting in 2021. Career statistics were calculated for each player based off of their recorded history. Certain players had switched teams during the 2021 season and as such, a column was added (team_change) to signify if they changed teams during the season. If the player switched to a team in the other league, their league ID was changed to “MLB” to indicate they played for both the National League and the American League, the two leagues of the MLB.

Since each player’s position corresponds to where they play in the defense, generally every player of any position will get an opportunity to bat in a season. And although many pitchers also bat, their batting statistics were significantly lower than players of other positions. As a result, all players listed as pitchers were removed from the analysis, as their high salaries are an outcome of a separate technique from batting, which would skew the data.

Career statistics were determined as the sum of all player data prior to and including the 2021 season. Batting average and career batting average were added as the ratio of hits to total at-bats. Once each individual was reduced to a single row, their real names were obtained from a table in the file titled People.csv, and salary information was retrieved from Spotrac (*MLB rankings*). The column “team” is the team that their salary info is from. Any players who did not match a salary from the top batters or pitchers were dropped from the set to get a better understanding of the top earners in the MLB. This was also a result of many of the bottom earners obtaining the same salary of $570,500. And as these players had varying numbers in their baseball statistics but the same salary, players making under $700,000 were removed from the dataset to avoid asymmetry in sample distribution. Finally, a simple random sample of 100 players were taken from the total of 231 for better viewing and inspection of the plots.

Through the data testing process, decisions were made to include new columns such as win percentage and win-loss ratio to represent the team they played on as a numerical value instead of team name, which was categorical. This was justified by the fact that every member of a team will have the same win percentage and win-loss ratios. As well, a player's draft pick was initially stored as the draft round and overall pick, but was changed to simply state whether they were drafted or not to account for free agents and undrafted players.

Statistical Modeling: Data Inspection & Response Transformation

The finalized dataset was loaded into R and a full model was made with salary as the response, and all other variables as predictors. Plots created to check model assumptions representing this full model can be seen in Figure 1. The required model assumptions of equal variance, linearity, and a mean of zero appear to be potentially violated. Due to this, both a log transformation of the response was tested, and a power transformation using the box-cox method.

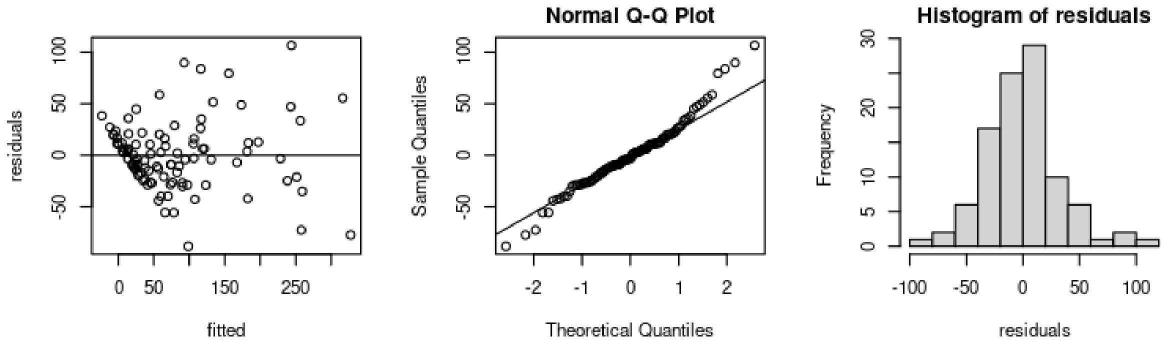

Figure 1*: Plots of untransformed response variable [y = salary ($100 000)]*

The ideal lambda found by the box-cox method was ~ 0.15, however a lambda of 0.2 was selected for simplicity as the difference between the sum of square residuals was small in scale. After applying the power transformation of 0.2 to the salaries, the plots were regenerated to check changes in model assumptions. The residuals vs. fitted values plot showed a slight improvement in linearity and equal variance. The Q-Q plot seemed to indicate a potential outlier as an influential point, with the rest of the points appearing to follow the normal line. Through investigation of this point, its Cook's distance was found to be approximately 0.441, relatively greater than the other points as shown in Figure 2. And although the Cook's distance was not greater than 1 as the rule of thumb suggests, the observation was removed for its influence on the response variable, relative to the other points. The Box-Cox power transformed model with the outlier removed that will be used for further analysis was then plotted in Figure 3.

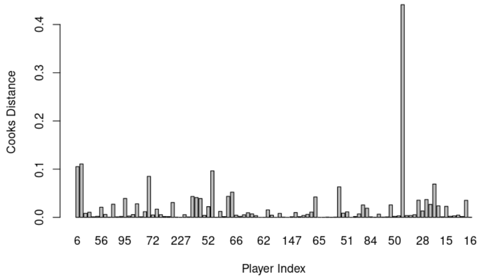

Figure 2*: Cook's distance bar graph*

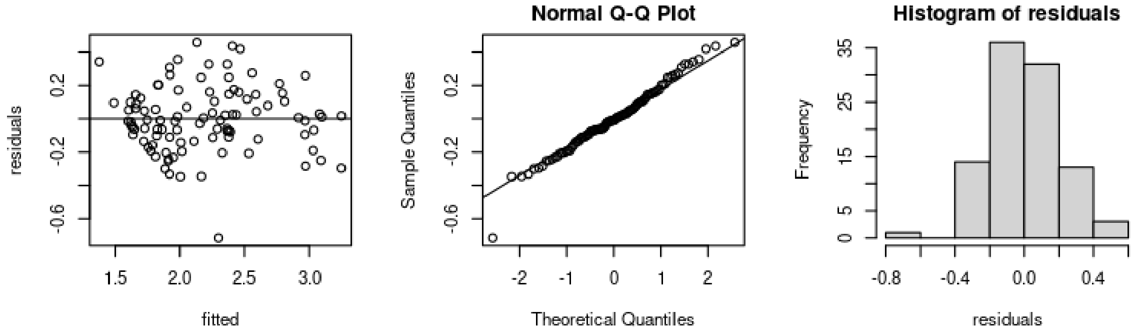

Figure 3*: Box-Cox power transformation of 0.2 on response with outlier removed*

Statistical Modeling: Model Selection

Model selection consisted of discrete processes under AIC and BIC to determine which variables would be removed or retained, as well as inspection of the results of the LASSO shrinkage method. To obtain minimized AIC values, forward selection, backward elimination, and stepwise selection algorithms gave proposed models and their adjusted coefficients of determination were collected. As well, to minimized BIC values, the R package "leaps" was used to give proposed models for forward, backward, and all-subset/exhaustive selection. Finally, the LASSO method was performed using the "glmnet" package, and two models were inspected; one model given by the error within 1 standard deviation of the minimum and the other generated by the lowest mean cross-validated error. The adjusted coefficients of determination were then compared to provide the final fitted model with coefficient estimates to predict a hitter's salary.

Results

A summary of the proposed models from AIC and BIC selection algorithms with their corresponding adjusted coefficients of determination were analyzed and provided in Table 1 and Table 2.

Method	Model	AIC	R^2_{adj}
Forward	salary ~ c_R + years_active + c_RBI + WP + AVG + H + c_BB + R	286.47	0.7924
Backward	salary ~ team_change + AB + R + RBI + A + E + c_G + c_H + c_R + c_BB + years_active + c_AVG + WLR	287.89	0.8041
Stepwise	salary ~ c_R + years_active + c_RBI + WP + AVG + H + c_BB + R	286.47	0.7924

Table 1: *Results of Forward, Backward, and Stepwise Model Selection with AIC*

Method	Model	BIC	R^2_{adj}
Forward	salary ~ c_R + c_RBI + years_active	-129.05	0.7673
Backward	salary ~ AB + c_H + c_BB + years_active	-123.17	0.7618
Exhaustive	salary ~ c_R + c_RBI + years_active	-129.05	0.7673

Table 2: *Results of Forward, Backward, and All Subset Model Selection with BIC*

The results of forward and stepwise selection proposed the same eight variables and same AIC value. Meanwhile, the backward elimination model provided thirteen variables, all different from those proposed in forward/stepwise selection and the associated AIC value was higher. Even though model selection using AIC would have the model of lower AIC to be optimal, the values of 286.47 and 287.89 were close enough to each other that their R^2_{adj} values were used instead for comparison. Therefore by AIC model selection, the backward elimination model was selected as the best proposed and would be used for comparison with other methods.

Much like with AIC, the forward selection algorithm proposed a model of the same three variables and BIC value as exhaustive/all-subset selection. The backward elimination model shared one variable with the forward/exhaustive model which was years active. The remaining three variables were different, and the model's BIC value was the greatest out of the three while also it provided the further coefficient of determination from 1. Since the resulting R^2_{adj} values were lower than the models proposed in the previous method, the AIC backward model remained the best for further comparison.

Next, the two models resulting from the LASSO method using the "glmnet" package were inspected. The first model with a λ = 8.974687 is the model with the largest lambda such that the error is within 1 standard deviation of the minimum. It utilized six non-zero coefficients, and a R^2_{adj} of 0.5308. The second model with a λ = 2.938806 is the model with the lowest mean cross-validated error. It used twelve non-zero coefficients and gave a R^2_{adj} of 0.6840. These results were inferior to the results gathered by AIC and BIC selection because of their smaller adjusted coefficients of determination, and therefore were not used in analysis.

Of the model selection methods, the AIC backward elimination model (Figure 4) was chosen as the best for predicting hitter salary because of its adjusted coefficient of determination's closeness to 1. To visualize this model's efficacy as it compares to the observed salaries, the fitted and observed points were plotted against each other (Figure 5). Although the proposed model did not appear to be perfect, the trends and clusters of the points for high versus low salaries were consistent in both the predicted and real values, making it a sufficient fitted model.

$$salary^{0.2} = 2.73799 + 0.13030 team\_change + 0.00088 AB - 0.00609 R + 0.00307 RBI - 0.00059 A + 0.01527 E - 0.00103 c\_G + 0.00119 c\_H + 0.00099 c\_R + 0.00110 c\_BB - 0.06394 years\_active - 4.04592 c\_AVG + 0.15672 WLR$$

Figure 4: *Backward Selection Selected Model*

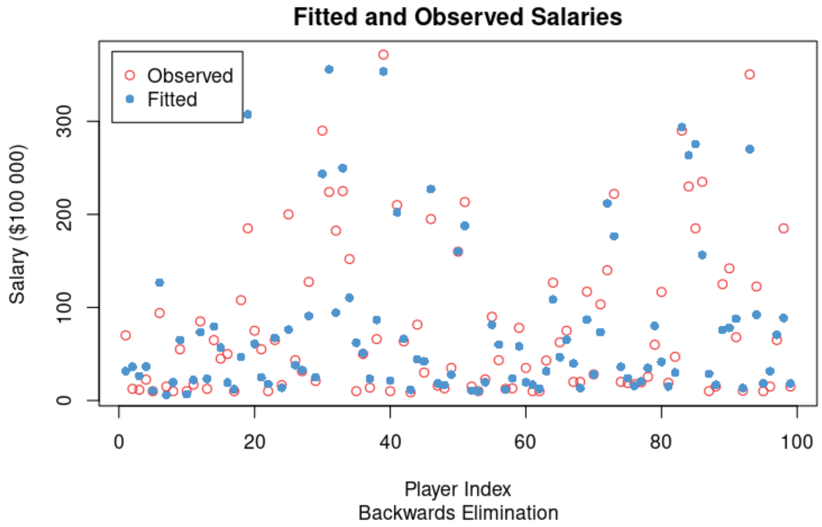

Figure 5: *Fitted and Observed Values for AIC Backward Elimination Model*

Discussion

Based on the initial hypothesis, there were a few unexpected results. Some statistics agreed with what was hypothesized such as runs and runs batted in having influence, as well as the abundance of career statistics as variables. Negative coefficients in the model may have been expected for some features such as player errors, but this was not the case for other predictors. Player errors had a positive coefficient, while the negative coefficients in the selected model include runs, assists, career games, years active, and career batting average. The unexpected coefficients of the defensive statistics (i.e. errors, assists) may suggest that they are not as important to a hitter's salary as their hitting statistics (i.e. runs, home runs, batting average, etc.). The variables years active and career games played make sense to have negative relationships with salary, as it suggests a player has passed his prime. Therefore, as a player ages or plays more seasons, their skill deteriorates, reflecting negatively on their next contract. The negative coefficients of runs and career batting average were surprising, but an increasing career batting average with decreasing salary may be a result of newer players having less plate appearances to compare their hits against. The model suggesting that as a player's runs in a season increase, their salary decreases may be a factor of runs not necessarily correlating to hitting performance. This can be explained by the fact that a majority of the runs a player obtains is ultimately scored by their teammates batting them in, excluding home runs, which are less common. Ultimately, the model is unexpected and may appear incorrect, but it can be explained through a greater understanding of the individual statistics of baseball.

Future Work

Although the results found have not been insignificant, there is clear room for improvement. This research focused on a subset of basic player statistics, and expanding these basic statistics to include further specifics such as positions played or left vs. right handed hitting may help narrow-down what really matters for a player's value. Additionally, many sports sites currently use an expanded set of more complex statistics, such as weighted averages which account for missed games. There is a strong likelihood that incorporating some of these more comprehensive statistics could improve the overall model accuracy. Research into how player salary negotiations work may also lead to unnoticed but important variables, such as specific arbitrators used in contract negotiations. Aspects of how a player's contract is mediated are hard to find information on and ultimately hard to measure. Without connection to an expert with a niche background, these were out of scope for this report. Re-completing a similar style analysis with pitchers and their associated statistics instead of hitters may also prove to be an interesting study.

References

Baseball savant MLB draft application. baseballsavant.com. (n.d.). Retrieved December 10, 2022, from https://baseballsavant.mlb.com/draft

Chadwickbureau. (n.d.). *Chadwickbureau/baseballdatabank: Development for Baseball Databank, an open data collection of historical baseball data*. GitHub. Retrieved December 10, 2022, from https://github.com/chadwickbureau/baseballdatabank

ESPN Internet Ventures. (n.d.). *2022 MLB stat leaders*. ESPN. Retrieved December 10, 2022, from https://www.espn.com/mlb/stats

MLB rankings. Spotrac.com. (n.d.). Retrieved December 10, 2022, from https://www.spotrac.com/mlb/rankings/

Appendix 1

Column	Data Type	Description	Column	Data Type	Description
weight	quantitative continuous	Weight in pounds	PO	quantitative discrete	Putouts
height	quantitative continuous	Height in inches	A	quantitative discrete	Assists
team_change	qualitative nominal	0=Unchanged 1=Changed	E	quantitative discrete	Errors
lgID	qualitative nominal	AL, NL, or MLB	AVG	quantitative continuous	Batting Average
years_active	quantitative continuous	Seasons played	c_G	quantitative discrete	Career games
salary	quantitative continuous	Salary in USD	c_AB	quantitative discrete	Career at bats
drafted	qualitative nominal	0=Unchanged 1=Changed	c_H	quantitative discrete	Career hits
G	quantitative discrete	Games played	c_HR	quantitative discrete	Career games played
AB	quantitative discrete	At bats	c_R	quantitative discrete	Career runs
R	quantitative discrete	Runs	c_RBI	quantitative discrete	Career runs batted in
H	quantitative discrete	Hits	c_BB	quantitative discrete	Career bases on balls
HR	quantitative discrete	Home runs	c_AVG	quantitative discrete	Career batting average
RBI	quantitative discrete	Runs batted in	WLR	quantitative continuous	Win loss ratio
BB	quantitative discrete	Bases on balls	WP	quantitative continuous	Win percentage