Impact of Home Runs on Runs Scored in the MLB

Trustin Vo

ECON 103A

San Jose State University

14 December 2020

**Abstract**

To better understand a highly valued metric of home runs in data analytics within the baseball world, this study analyzes the relationship between home runs on runs scored in Major League Baseball (MLB) using linear regression. Alongside home runs, additional variables such as hits, walks, and team league were factored into a multivariate regression. Using data from the 2019 MLB regular season provided by Baseball Reference, I hypothesize that home runs play the biggest factor on runs scored. The results of the regression models agree with this hypothesis.

**Introduction**

After the popularization of the Michael Lewis biographical book, *Moneyball* (2004), data analytics within the sports world rose to prominence after a 2011 film was adapted. The book chronicles Oakland A’s manager, Billy Beane was able to find production in undervalued players to navigate around a low budget in the 2002 MLB regular season. While other teams paid high dollar value for highly valued statistics, such as batting average, Beane was able to use linear regression to value a new metric, which is referred to on-base percentage. This team production was highlighted by a 20-game winning streak into the playoffs, finishing first place. This method of empirical analysis is dubbed “sabermetrics”, which refers to the application of mathematical and statistical analysis in baseball.

The goal of this study is to determine the relationship between a highly valued statistic in baseball, which is runs scored (which can be thought of as “points scored” in other sports), with a highly valued statistic -- home runs. Naturally, this association is a generally accepted relationship among pop culture, as home runs are one of the most exciting and better-known assets of the sport of baseball. However, this study is aimed to numerically value this statistic. If the analysis reveals there is a lack of association between the two, other factors will be tested to observe what constitutes “runs scored” in baseball.

This study relates to an economic journal published by Hakes and Sauer (2006), which uses econometric tools to confirm Lewis’ story. Hakes and Sauer study revealed that home run hitters were heavily more valued than other players, earning “$3-$4 million more than the average player” (p. 176). They state that on-base percentage was an undervalued statistic from 2000-2004, which was the period before the publication of *Moneyball*, verifying its relevance. Their analysis was able to conclude that the labor market in baseball was not efficient prior to *Moneyball*, due to overvaluing home runs and other offensive statistics. This study will evaluate why home runs hold that value in the baseball viewer’s minds.

**Data**

The data in this study was retrieved from the 2019 MLB season, provided by Baseball Reference. Baseball Reference is a renowned resource for baseball statistics within the sports industry. The data was downloaded into an Excel file, which was where the regression was also ran. The variables within the study are defined in the following table, Table 1. A descriptive statistic summary follows in Table 2.

Table 1: Variables defined

| Variable | Definition |
| --- | --- |
| runs | The total amount of runs scored by offenses league-wide in 2019 |
| HR | The total amount of home runs hit by hitters league-wide in 2019 |
| h | The total amount of hits (minus HR) by hitters league-wide in 2019 |
| bb | The total amount of walks drawn league-wide in 2019 |

Table 2: Summary statistics

| Variable | Observations | Mean | Standard Deviation | Minimum | Maximum |
| --- | --- | --- | --- | --- | --- |
| runs | 30 | 782.23 | 16.44 | 582 | 943 |
| HR | 30 | 225.87 | 7.61 | 146 | 307 |
| h | 30 | 1175.43 | 12.64 | 1052 | 1334 |
| bb | 30 | 529.83 | 13.08 | 378 | 645 |

Because there are 30 teams in the league, 30 observations were conducted. Each team scored from the range of 582 to 943 runs scored, with a mean of 782.23, with notable variation. The amount of home runs by each team ranged from 146 to 307 home runs, yielding a mean of 225.87 and slight variation. Since the purpose of this study was to determine how home runs affect runs scored, runs scored must be the dependent variable in the model.

The simple regression equation that this data applies to is as follows:

runs = 𝛽0 + 𝛽1*hr* + *u*

runs = 368.49 + 1.832*hr* + *u*

The regression output can be viewed in the appendix, under Appendix 1.

Additional variables were included to conduct a multivariate regression. Key offensive statistics, such as hits and walks are also valuable in scoring runs. Hits and walks were added to the equation to test for its effect on runs scored. Since home runs are considered hits, home runs were subtracted from hits to avoid double counting and those values were assigned the *h* variable. From 30 observations from around the league, each team ranged from 1052 to 1334 hits (minus home runs), averaging 1175.43 non-home run hits, with a notable variation of 12.64. Walks ranged from 378 to 645, with an average of 529.83 and a standard deviation of 13.08. The regression output can be viewed in the appendix, under Appendix 2.

The multivariate regression equation that accounts for hits and walks is as follows:

runs = 𝛽0 + 𝛽1*hr* + 𝛽2*h* + 𝛽3*bb*+ *u*

runs = -4663.388 + 1.451*hr* + 0.608*h* + 0.39*bb* + *u*

The addition of two variables into the regression reveals a minimal change in the original variable, home runs. From this equation, we can conclude that home runs are the most significant factor in calculating runs scored. The equation states that a home run leads to 1.451 runs scored, on average. Hits and walks play a smaller role, leading to an average of 0.608 and 0.39 runs scored, respectively. It can then be deduced that on average, home runs provide 0.843 more runs scored than hits and 1.061 more runs scored than walks. These findings align with the original hypothesis that home runs in fact are the most significant factor in accounting for runs scored in the MLB league in 2019.

To test for significance, the multiple regression model yielded the t-statistic, p-value, and confidence levels for the independent variables are listed in Table 3:

Table 3:

| Independent Variable | Standard Error | t-statistic | p-value | Confidence Interval |
| --- | --- | --- | --- | --- |
| hr | 0.178 | 4.718 | 0.00 | 0.475 - 1.21 |
| h | 0.065 | 9.268 | 0.00 | 0.473 - 0.743 |
| bb | 0.096 | 4.071 | 0.00 | 0.193 - 0.586 |

The source of these values can be seen under Appendix 3.

Additionally, as a result of the interpretation of the p-value, the variables for home runs, hits, and walks are deemed significant. Their significance is backed by low p-values. The p-values indicate that there is only a 5% chance the variables lie outside of the confidence intervals. The question may arise, how correlated are these variables with runs scored, exactly? The r-squared value provided by the multivariate regression output seen in Appendix 2 indicates the r-squared value to be 0.938. This indicates high correlation. With such high t-statistics, there is indication that there is significant difference between the variables. To test for multicollinearity, standard errors and coefficients were analyzed. Because the standard error for the coefficients were low, there is no evidence pointing towards multicollinearity. To test for heteroskedasticity, the trendlines in correlation between the independent variables and dependent variables can be observed. In Chart 1-3, the points on the scatterplot are closely aligned to the trendline. Because there is high correlation, it is assumed that the data is highly heteroskedastic.

Chart 1:

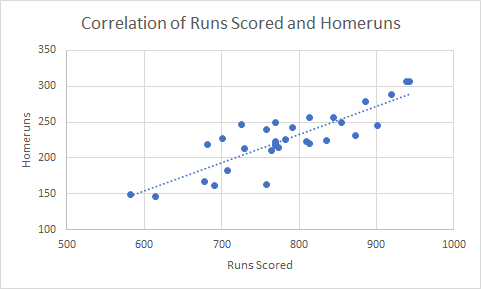


Chart 2:

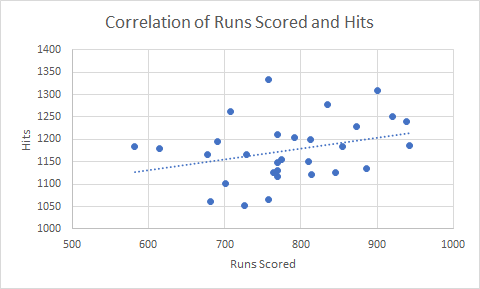
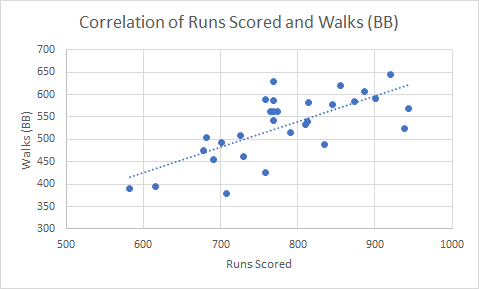


Chart 3:

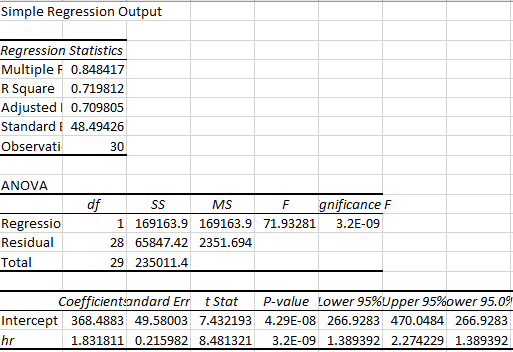


**Conclusion**

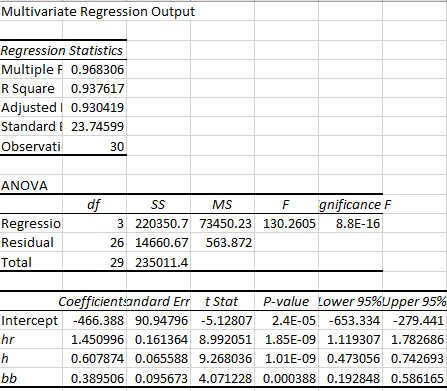
With the goal of testing the relationship between runs scored and home runs in the econometric scope, two regression analyses were performed. To test the association, both a simple regression analysis and a multivariate regression analysis revealed the impact of home runs, hits, and walks on runs scored in Major League Baseball. Using data from 2019 provided by Baseball Reference, an econometric model was formed to test this relationship. As a result of regression analyses, this research concludes that home runs in fact play the biggest role in producing runs scored in Major League Baseball. This study originally hypothesized that home runs will play the biggest factor in runs scored compared to other variables, such as hits and walks. The empirical analysis confirmed this hypothesis, with variables considered to be statistically significant. With a high r-squared value, the independent variables proved to be strongly correlated with the dependent variable. While other analytical methods are used to value under-valued statistics in baseball, home runs prove to be the most significant statistics in offensive scoring in the MLB.

**Appendix**

Appendix 1:



Appendix 2:



Bibliography

2019 Major League Baseball Season Summary. (n.d.). Retrieved December 10, 2020, from https://www.baseball-reference.com/leagues/MLB/2019.shtml

Hakes, J. K., & Sauer, R. D. (2006). An Economic Evaluation of the Moneyball Hypothesis. *Journal of Economic Perspectives,* *20*(3), 173-185. doi:10.1257/jep.20.3.173