Tom Bash and Conor Keating

Methods of Data Science

Dr. Dumnich

May 4, 2021

Summary of Findings

The goal of this study was to find the most influential variables and baseball statistics when trying to predict a MLB players' first year arbitration salary. After conducting multi linear regression tests and analyzing the p-values, R values, R-Squared values, and the equations of the lines of best fit, we were able to get a good idea of which variables impact salary the most. The variables that were statistically significant and had a p-value lower than .05 were:

* Batting Average (BA) (pval: 2.91674\*10^(-6))
* On base % (OBP) (pval: 2.65889\*10^(-7))
* Home runs (HR) (pval: 5.1998\*10^(-16))
* Runs Batted In (RBI) (pval: 4.19675\*10^(-16))
* Runs (R) (pval: 3.20004\*10^(-15))
* Slugging % (SLG) (pval: 5.07875\*10^(-10))
* On Base + Slugging (OPS) (pval: 3.61482\*10^(-11))
* Plate appearances (PA) (pval: 6.06771\*10^(-11))

The variables that were not statistically significant were:

* Stolen Bases (SB) (pval: 0.106251966)
* Strikeouts Percentage (Kper/K%) (pval: 0.48084081)
* Walks Percentage (BBper/BB%) (pval: 0.007878414)
* Defensive Runs Saved (DRS) (pval: 0.713261653)

At the conclusion of this study, we have found which variables are the most influential when determining first year player arbitration salary. This study can be used by players to see which stats can give them a bigger pay day. This study can also be used by General Managers who can now better estimate what a player should be paid based on their hitting and fielding statistics.