When assessing the baseball data, this study hoped to identify whether a baseball player’s offensive stat such as hits, home runs, and RBIs significantly impacted salary. This study also wanted to explore whether the number of years in a league impacts performance and if a player’s league, American versus National, impacts salary.

The regression analysis using statsmodels showed that the model only explained 22.1% of the variance in salary, meaning that there are other important factors besides hits, home run, and RBIs that impact salary. Hits and RBIs were positively significant, while home runs were negatively insignificant. Scikit-learn’s regression results showed slightly different results, in that, all three variables had a positive impact on salary. In total, these findings indicate that altogether these three variables contribute to a player’s salary; however, they are not the only variables that influence salary. The PMF and CDF analysis results indicated that players with more years of experience tend to perform better in terms of hits and home runs. The permutation test analysis showed that the impact league had on salary was statistically insignificant.

A major factor that was missing during the analysis was the lack of variables used. Based on the results, it showed that salary is influenced by more than just hits, home runs, and RBIs. The analysis could also have assessed player awards, contract negotiations, contract length, endorsements, and defensive stats. The regression model also failed to take into account positional differences, which could have played a huge part in salary earnings.

An incorrect assumption was that offensive stats alone such as hits, home runs, and RBIs could predict salary with high precision. The low R-squared value suggests that salary is influenced by several other key factors. Another assumption was that each offensive stat contributes independently to salary whereas the results showed that there might be multicollinearity issues which likely skewed the findings. The study also assumed that all players are evaluated equally despite positional differences, overlooking the team structure.

One challenge was understanding how home runs had a negative insignificant coefficient. This could be due to multicollinearity issues, where home runs are highly correlated with the other two variables (hits and RBIs), making it hard to tell their real impact in the analysis. Based on the results of the regression analysis, future studies should examine how statistical libraries, statsmodels and sciki-learn, handle the data in different ways and determine which regression model would have been the most appropriate to use.

In summation, hits and RBIs are stronger salary predictors than home runs, and salary is influenced by additional variables not included in this analysis. The study also showed that veteran players tend to score more hits and home runs than new players. Future research should examine other variables that could have a significant impact on salary as well as consider other modeling techniques such as decision tree or random forest. Future studies should also use VIF analysis to identify multicollinearity in the regression models.