THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

University of North Carolina at Chapel Hill

Department of Statistics and Operations Research

STOR 664: Project – Phase I

Students:

Mark Cahill, Jack McPherson, Gabriel Sargent, Hanieh Jamshidian

Instructor: Professor Daniel Kessler

Fall 2025

1. Research Question

We will study a linear model for predicting the salaries of Major League Baseball (MLB) players as a function of their hitting statistics, like batting average, number of home runs, and number of runs batted in. We will test whether each of the hitting statistics has an impact on salary (we expect they will), so our null hypotheses will be of the form:

$$H_0: \beta_i = 0$$

We will also study the relative importance of the statistics (e.g. does batting average have a greater impact on salary than number of RBIs?), so we will consider null hypotheses of the form:

$$H_0: \beta_i > \beta_j$$

And will compute partial R^2 statistics in order to see how much salary variation is explained by each predictor. If time permits, we may then repeat this analysis on individual teams to see how the relationship between batting statistics and salary varies across teams.

2. Dataset Description

The dataset used is from Sean Lahman, and we ended up combining multiple datasets of his to create one large dataset we will operate on. Specifically, we used his datasets on salaries, batting, pitching, and his master dataset to track each player's earnings by year, in-game stats, and tie them to their name. Each dataset uses the identifying variable "playerID" which is matched to a real name in the master dataset. This way, we can identify individual players that stand out.

We are specifically working from the time period of 1985-2015, and in this period, certain teams had pitchers also hitting. We decided, since pitchers were typically not paid for their hitting stats, that their (usually) poor performances in batting were not worth including in our dataset.

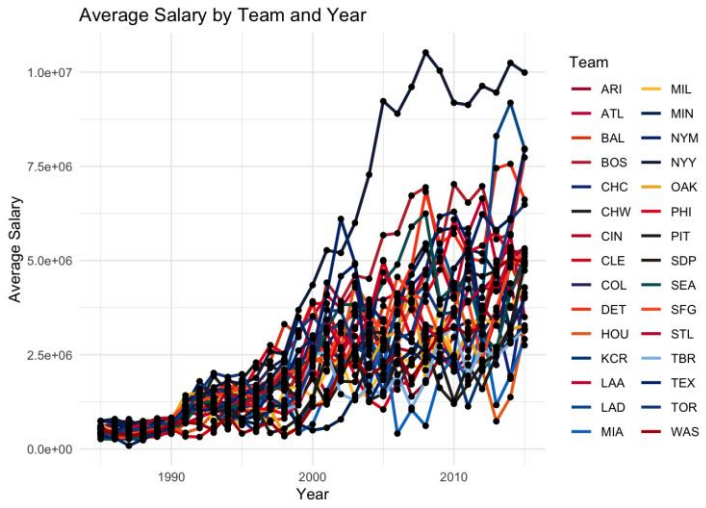
3. Background & Prior Work

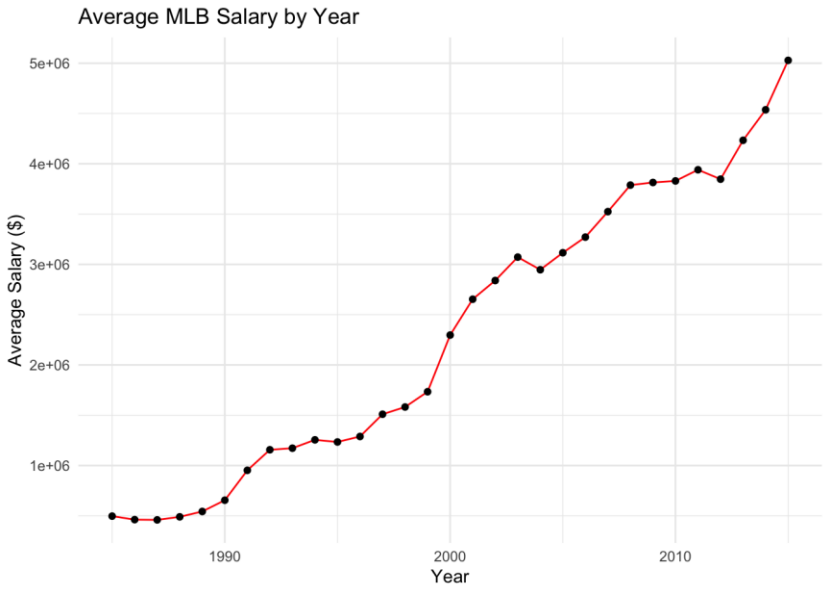
Many studies have examined how baseball players' performance is related to their salaries. One of the earliest examples is by Scully (1974), who used MLB performance and salary data and applied regression models to estimate how much value players create for their teams. His results showed that important hitting and pitching statistics, as well as experience, have a strong effect on salary [1]. Later, Yilmaz and Chatterjee (2003) looked at the same topic but also considered team goals. They found that player performance matters, but salaries can also change depending on what the team is trying to achieve [2]. More recently, Magel and Hoffman (2015) used data from 2010–2012 and built stepwise regression models. They showed that career totals; such as total hits, home runs, strikeouts, and saves; predict salary better than only using one season of statistics [3]. Another study by DeBrock, Hendricks, and Koenker (2004) analyzed more than ten years of MLB data and

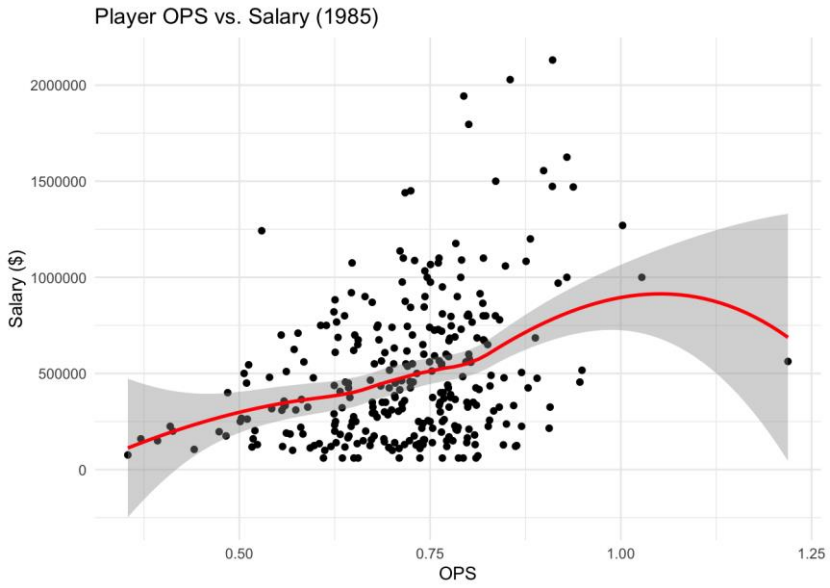
focused on team-level salary distribution. They found that when salary differences inside a team are large, team performance usually becomes worse [4].

Several practical projects on Kaggle also use the Hitters dataset [5, 6] or the Baseball Databank [7] to explore salary prediction. These projects mainly use linear regression, decision trees, and exploratory analysis. Their results are like academic studies, i.e., offensive performance measures such as home runs, hits, batting average, and slugging percentage, especially when combined with career totals, have the strongest relationship with player salary. Overall, the existing literature shows that performance statistics explain a large part of salary variation; career-long performance is often more useful than single-year data, and sometimes team-level factors also play a role in salary decisions.

4. Exploratory Figure




Average MLB Salary by Year
5e+06
4e+06
3e+06
2e+06
1e+06
Average Salary ($)
1990
2000
2010
Year


Player OPS vs. Salary (1985)
2000000
1500000
1000000
500000
0
Salary ($)
0.50
0.75
1.00
1.25
OPS

5. Data Concerns & Wrangling

The data itself seems rather thorough; however, the only concern is ensuring all pitchers are removed from the dataset. This would not normally be a concern since using the anti_join() function would accomplish this, but in certain situations, teams may have one of their position players pitching if the game is very lopsided. What this does is it creates stats for said position player as a pitcher, and removing this player from the dataset would remove him as a hitter as well even if that is his priority.

To avoid this, we restricted all possible hitters to only those with at least 50 AB's in a season, and we removed only pitchers with at least 18 innings pitched. This accomplishes a few things: it first hopefully reduces the chance of accidentally removing a position player who pitched a couple of times (since actively searching through every game to find examples of this would be very tedious), and secondly it removes players who only played a couple of games and were otherwise not major factors for the whole season. The focus is intended to be on players who played a majority of games and theoretically had a season to accumulate a solid amount of counting stats. The lone downside is this may have removed what would be everyday players who battled injuries which caused them to miss significant chunks of certain seasons. However, this could be seen as a positive since if said player could not meet the counting stats worth a major salary, then said player could show up as an outlier in our dataset.

6. Analysis Plan

Our plan for this data is to develop a linear model to predict player salaries based on various batting statistics. The first step will be to ensure that our variables are normally distributed and apply transformations if necessary. We will then fit a linear model to try to predict salary. Since our data set has a lot of potential predictors, we will then try to narrow down the variables in our linear model. Finally, we will test the relative difference between the regression coefficients to identify the variable that has the greatest effect on player salary.

7. References

[1] Scully, G. W. (1974). Pay and Performance in Major League Baseball. The American Economic Review, 64(6), 915–930. http://www.jstor.org/stable/1815242

[2] Yilmaz, M. R., & Chatterjee, S. (2003). Salaries, Performance, And Owners' Goals In Major League Baseball: A View Through Data. Journal of Managerial Issues, 15(2), 243–255. http://www.jstor.org/stable/40604428

[3] Magel, R., & Hoffman, M. (2015). Predicting salaries of major league baseball players. International Journal of Sports Science, 5(2), 51-58.

[4] Debrock, L., Hendricks, W., & Koenker, R. (2004). Pay and Performance: The Impact of Salary Distribution on Firm-Level Outcomes in Baseball: The Impact of Salary Distribution on Firm-Level Outcomes in Baseball. Journal of Sports Economics, 5(3), 243-261. https://doi.org/10.1177/1527002503259337 (Original work published 2004)

[5] https://www.kaggle.com/code/nihandincer/hitters-baseball-data?utm_source=chatgpt.com

[6] https://www.kaggle.com/code/ozlemilgun/salary-prediction-with-hitters-data-set#Summary

[7] https://www.kaggle.com/code/samfenske/ml-predicting-mlb-salaries