THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

University of North Carolina at Chapel Hill

Department of Statistics and Operations Research

STOR 664: Final Project

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:

$$H_0\colon \beta_i = 0$$

We will also study the coefficients in batches, such as

$$H_0\colon \beta_i = 0,\ i = 1, \dots, p$$

as well as some subset:

$$H_0\colon \beta_i = 0,\ i \in \mathrm{I} \subseteq \{1, \dots, p\}$$

We will compute partial R^2 statistics in order to see how much salary variation is explained by each predictor, as well as using VIF analysis for model selection.

2. Dataset Description

The data used in this project comes from Sean Lahman's Baseball Database, which is available on Kaggle. For our analysis, we combined information from four different datasets: Salaries, Master, Batting, and Pitching. Each of these datasets uses the same identifying variable, playerID, which allows us to merge the tables and connect every player's salary, performance statistics, and personal information.

From the Salaries dataset, we use each player's reported annual salary together with yearID, teamID, and league (lgID). From the Batting dataset, we include the main hitting statistics needed for our model, such as AB (at-bats), H (hits), HR (home runs), RBI, BB (walks), and SO (strikeouts). Using these, we also compute important performance measures including batting average (Avg), slugging percentage (Slg), on-base percentage (OBP), and OPS. The Master dataset provides players' first and last names, and the Pitching dataset is used only to identify pitchers so we can exclude them, since pitchers are not typically paid based on hitting performance. Our final merged dataset covers the years 1985–2015 and contains approximately 12120 rows and 30 variables.

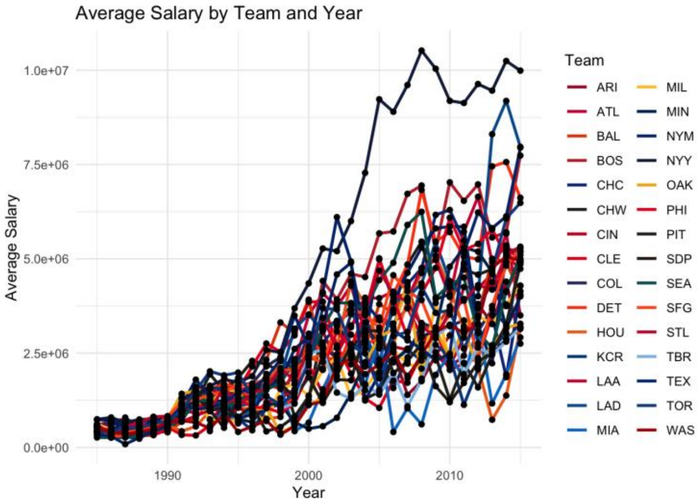
3. Background & Prior Work

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.

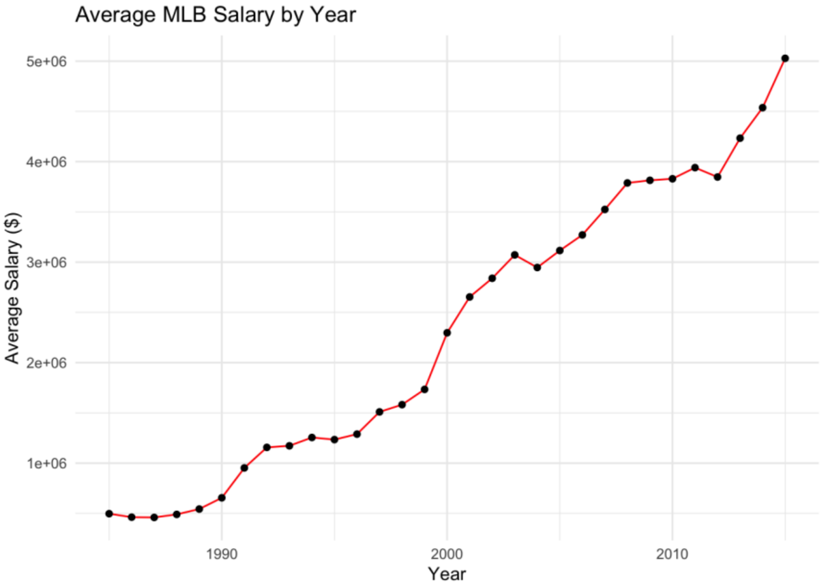
4. Exploratory Figure

The first plot shows how the average salary for each MLB team has changed over the years. Salaries stayed low until the early 1990s, then increased sharply, especially from 1995 to 2010. In later years, some teams show much higher average salaries than others, which creates a wider gap between high-paying and low-paying teams. Overall, the figure shows strong long-term salary growth in MLB and increasing differences in spending across teams.

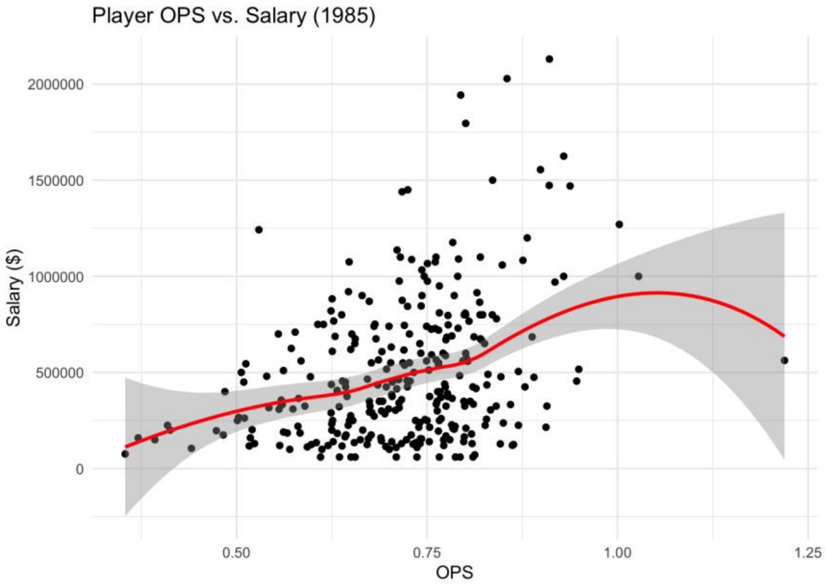


The second figure shows the average MLB salary for all players by year. The trend is clearly upward; salaries stay low until the late 1980s and then begin rising quickly through the 1990s and

2000s. There are small fluctuations in some years, but the overall pattern is steady growth. By the end of the period, the average salary is several times higher than at the beginning. This plot illustrates how MLB salaries have increased strongly over time.



The third plot shows the relationship between OPS (On-base Plus Slugging) and player salary for the year 1985. Each point represents one player. The general trend is upward; players with higher OPS values tend to have higher salaries. The red curve shows a smoothed line, and it also rises overall, which supports the idea that better offensive performance is linked to higher pay. The spread of points shows some variation; not all high-OPS players earn high salaries, but the positive pattern is still clear. This figure suggests that OPS can be an important predictor of salary in 1985.



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. Data Preprocessing and Cleaning

Before building our model, we performed several preprocessing steps to make sure the data we use is clean, consistent, and meaningful for the analysis. We first filtered the dataset to include only observations from 1985 and later and players with at least 50 at-bats (AB ≥ 50); this removes players who have very limited playing time, since their salary and performance would not be comparable to regular players. We also added several common rate variables: batting average, on-base percentage, slugging, and on-base plus slugging (OPS). During 1985–2015, some pitchers still had batting statistics. As we mentioned earlier, because pitchers are usually not paid based on their hitting, their batting numbers can be very low and distort the analysis. To avoid this, we removed cases where position players were pitching. We assumed that a position player would not pitch for more than 54 outs (18 innings) in a season.

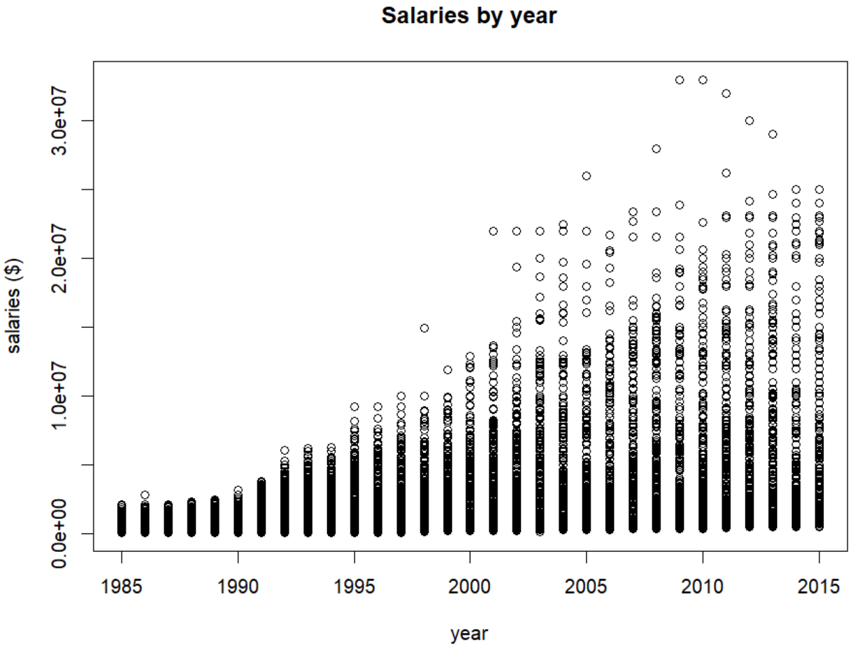
One challenge in our dataset is that MLB salaries increased over the years, as shown in the previous plots. Comparing raw salaries across different decades is misleading. To handle this, we scaled each player's salary by the maximum salary of that same year:

$$Scaled\ salary\ = \frac{Salary}{Max\ salary\ of\ that\ year}$$

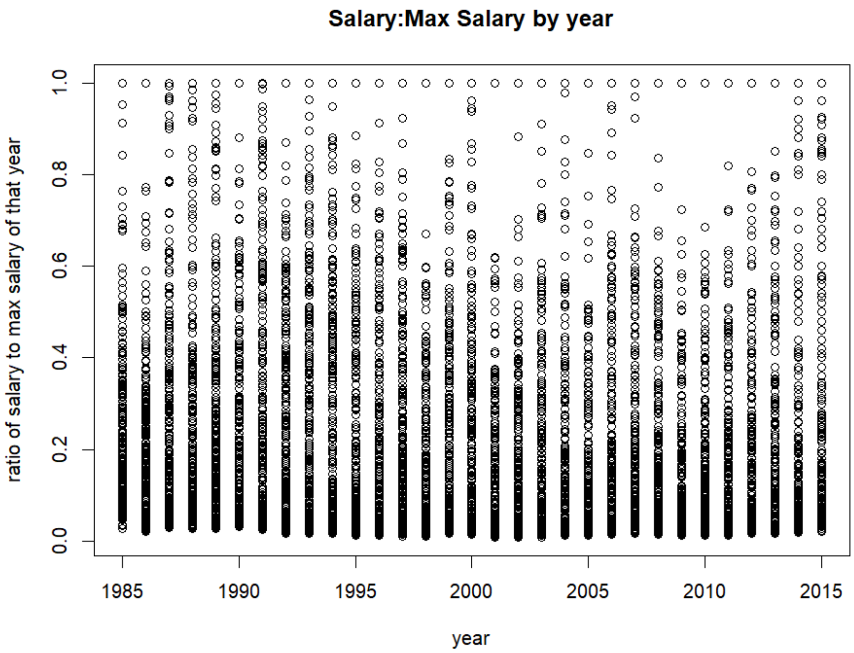
This transformation puts all salaries on the same scale, with the highest-paid player in any year having a value of 1. Even after scaling, the new response variable is still highly skewed toward zero. To improve the distribution, we applied a natural log transformation:

$$Log\left(\frac{Salary}{Max\ salary\ of\ that\ year}\right)$$

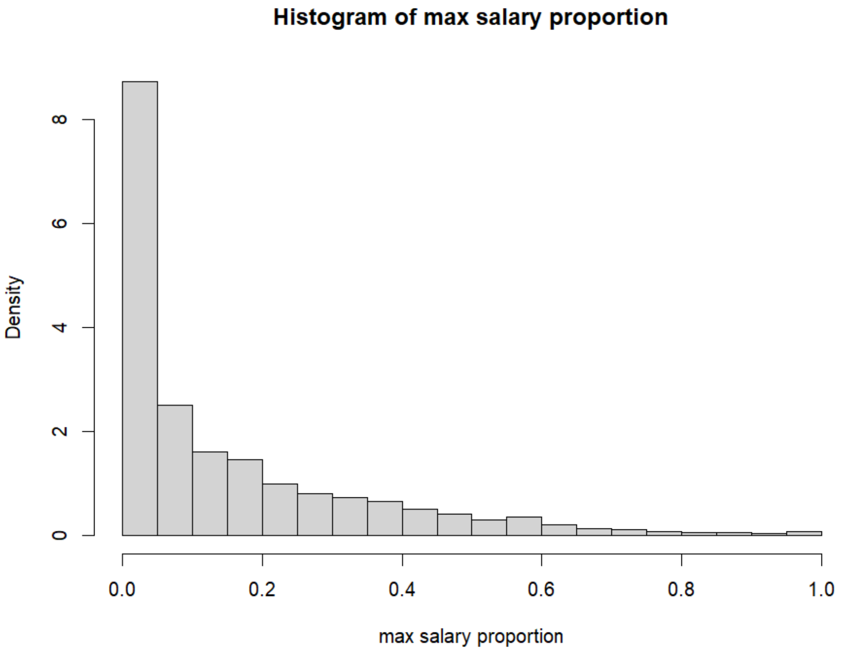
Below, we include several plots that help show why these cleaning and transformation steps were necessary.



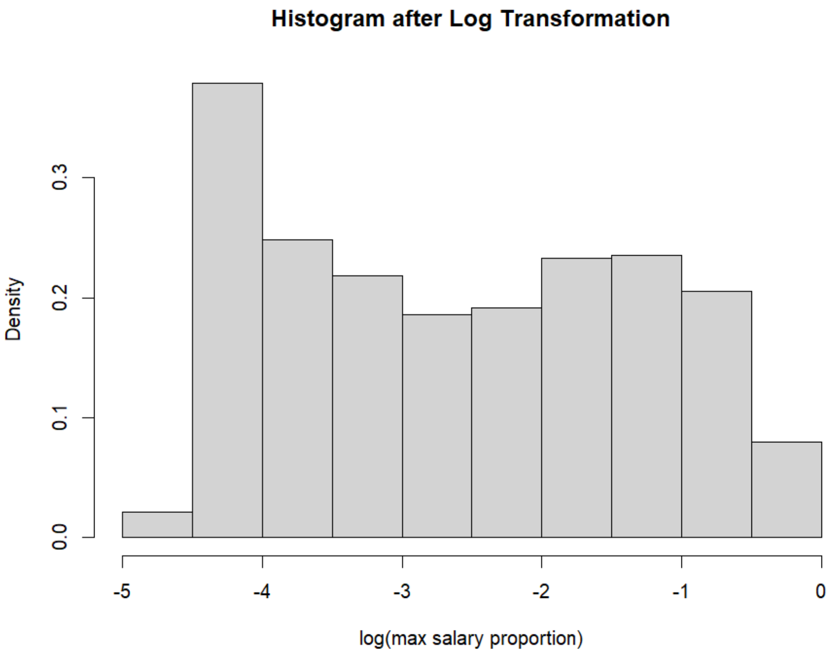
The above plot shows that salaries increase sharply over time, especially after the 1990s. It illustrates why comparing raw salaries across years is difficult.



After scaling the max salary of each year, all values fall between 0 and 1. This removes the long-term inflation trend, making salaries from different years more comparable.



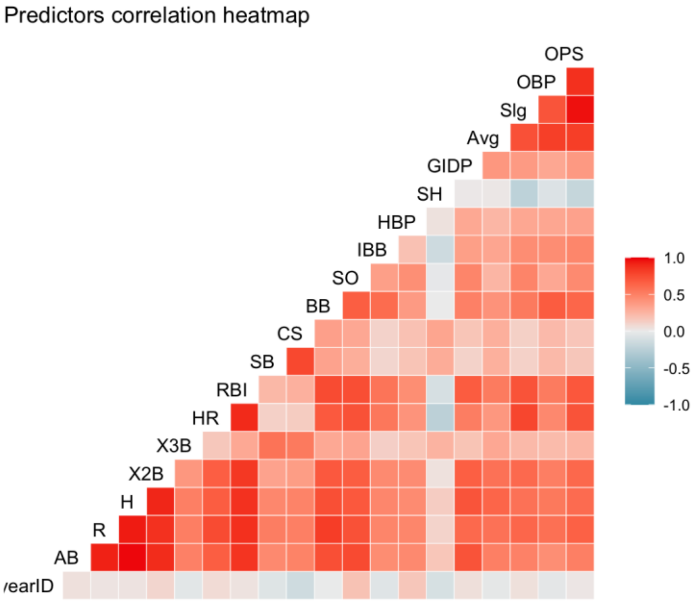
The distribution of scaled salaries is extremely right skewed, with many players earning a small fraction of the top salary. This motivated us to perform a log transformation.



After taking the natural log of the scaled salary, the distribution becomes much more balanced and closer to a normal shape. This is more appropriate for linear regression.

7. Analysis

Perhaps unsurprisingly, there is a fair amount of multicollinearity among our predictors. To avoid this, we used the following correlation heatmap to identify pairs of highly correlated predictors so we could avoid using them both in our analysis.



Based on the above heatmap, it is clear to see that generally the predictors are positively correlated. This is not a surprise, nor something we were overly concerned about seeing; however, we were looking to eliminate instances of extreme correlation. An example of this is in the case of OPS, which by definition is On Base Percentage (OBP) plus Slugging (Slg). Thus, we felt it made sense to only use one of the decimal-based statistics we added to the dataset. We ultimately chose OPS since it accounts for the most outcomes of those four statistics, and it is an increasingly popular tool to measure a player's success because of its added weighting of extra base hits.

From here, we looked at other highly correlated pairs and chose one of each to add to our model. We chose RBIs instead of Homeruns since Homeruns were already in-part measured in OPS. We also added Strikeouts (SO), Walks (BB), and Intentional Walks (IBB), and Stolen Bases (SB). We did not include Sacrifice Hits (SH) or Groundout into Double Plays (GIDP) since those are more niche and not really reflective of a player who would get paid more. Intentional Walks are also situational, but it was worth entertaining the idea that a batter that pitchers were more frequently afraid to pitch would be paid more.

We used VIF on our 6 predictors and generated the following values.

Predictor	VIF
Walks	3.307605
Strike Outs	2.648953
Stolen Bases	1.162702
RBIs	4.269219
OPS	2.164717
IBB	1.766000

Each of the predictors has corresponding values less than 5, which means the variables here are not strongly correlated with each other.

We also studied R^2 statistics to see which batting statistics are most strongly associated with salary. Specifically, for each statistic (walks, strike outs, stolen bases, runs batted in, on base plus slugging, intentional walks), we fit a linear model predicting rewards based solely on that statistic. These are the R^2 values for the linear models corresponding to each statistic:

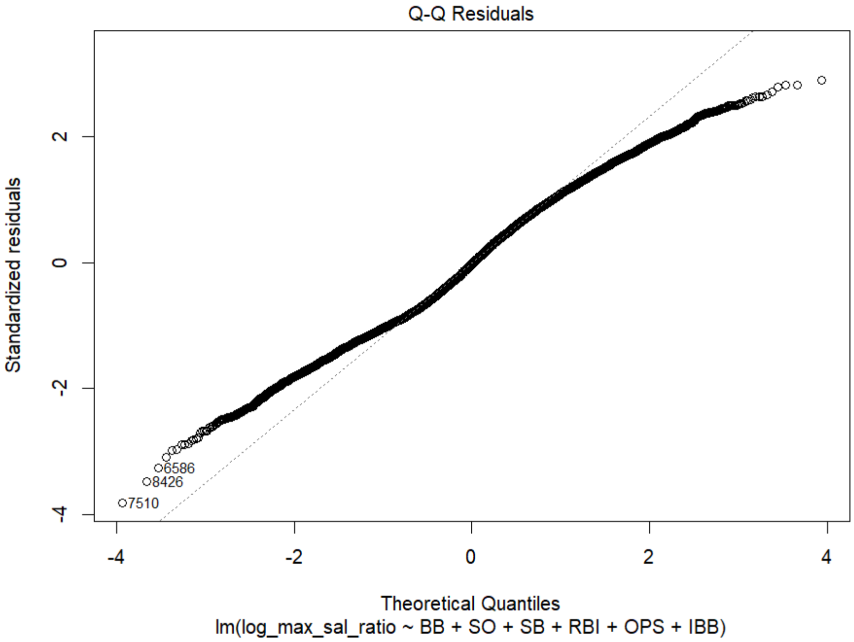
Predictor	R^2
Walks	0.376
Strike Outs	0.511
Stolen Bases	0.202
RBIs	0.394
OPS	0.740
IBB	0.141

These results suggest that variation in OPS (on base plus slugging percentage) explains the greatest amount of variation in salary, whereas variation in intentional walks explains the least. This finding makes sense. Intentional walks are somewhat rare and are mostly applied to elite hitters, so it is not a very meaningful statistic since most players have none. OPS, however, does not suffer from this sparseness, and it captures both the frequency (on base percentage) and extent (slugging) of a player's offensive contributions.

Finally, we fit a model with the selected predictors regressed onto the log of the scaled salary. We obtained the following model, as well as the test statistics and p-values of t-test for each individual predictor.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.67704	0.076912	-34.8064	4.39E-253
BB	0.01592	0.000757	21.02161	2.16E-96
SO	-0.00923	0.000457	-20.1896	3.51E-89
SB	0.002168	0.001086	1.996964	0.045851
RBI	0.017535	0.000683	25.67517	1.30E-141
OPS	-1.03185	0.121468	-8.49482	2.21E-17
IBB	0.024196	0.003168	7.638144	2.37E-14

Below is the qq plot of the residuals. As we can see, there is room to improve the model, so the residuals would be closer to normality (indicated by the 45-degree line in the plot).



First let us consider the hypothesis:

$$H_0\colon\ \beta_i = 0\ \forall\ i = 1, \dots, p$$

This gives a F-test p-value of 0.000, which indicates that we can reject the idea that none of them have an effect. Now consider the hypothesis:

$$H_0\colon\ \beta_i = 0\ \forall\ i \in \{BB,\ SO,\ OPS\}$$

This has a p-value of 1.10e-137, so we can also reject this null hypothesis.

8. Challenges during Analysis

The primary issue in our analysis is dealing with the serial correlation in the salaries. MLB salaries have skyrocketed since 1985. This is partially due to natural inflation, but even accounting for that players have just gotten paid more. Ideally, this would lend itself to some time series analysis. However, we would have to track each player as their own time series, which simply is not feasible when you consider that we have a total of 2149 players in our dataset. We think that scaling by the max salary by year is a good compromise, but there is still going to be issues with autocorrelation that would require other methods to fix. We also did not consider transformations of the predictor variables, though this was mostly due to time constraints. A future analysis of this subject might be better off using a non-parametric model such as loess regression.

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