Factors in Determination of Player Salaries and Win Percentages

Matthew Coleman and Stephen Lantin

June 2019

Abstract

We explore the relationship between the salaries of baseball players and game performance using Lahman's Baseball Database. Using data taken in the years 1990 to 2000, we determine that in-game performance is not a good predictor of salaries (30% correct with a tolerance of $100,000). Additionally, we assess team performance and succeed in correctly predicting 90% of win percentage with a tolerance of 5%, using the ratio of runs to earned runs. We anticipate this information to be useful to baseball team executives who aim to maintain the financial health of their franchises.

1 Introduction

In 2001, the Oakland Athletics suffered a devastating loss to the New York Yankees at the American League Division Series. Foreseeing star players leaving the team and armed with but a limited recruiting budget, Athletics general manager Billy Beane teamed up with economics graduate Peter Brand to rebuild the franchise. Such was the birth of sabermetrics, a method of assessing baseball player value based on past game performance. Using this recruiting strategy, the Athletics went on to achieve the longest winning streak in American League history, a story made famous in the film "Moneyball." Nearly two decades after the A's historic season, sabermetrics continues to bring large returns for baseball teams all across America and the world. Given its direct impact on in-game success, however, baseball team executives have in recent years seen another use for the science–optimizing profit.

To improve the financial health of a baseball franchise, analysts look to major contributors of revenue and expenses for improvement. In Major League and Minor League Baseball, for example, player salaries accounted for 54.2% of league revenue in 2018 [1]. While there are non-performance factors that weigh into the salary-determining process, franchise executives may find it difficult to determine if they are overcompensating their players. As budding data scientists, we can draw insight from raw data to aid such business decisions.

2 Questions of Interest

If the goal is to make a baseball franchise a healthy investment, it follows that executives should provide salary compensation according more to in-game performance and less on other factors. As such, we look at past data to answer the question, "What, historically, have been suitable in-game performance predictors of player salaries?" To answer this question, we are going to explore the relationship between salaries and many different batting statistics, and fit a regression to the data. We are going to fit a model based on solely hitting statistics, one based on all statistics, and then another model based on non-hitting statistics. The model containing non-hitting statistics will contain measures such as intentional walks, stolen bases, and strikeouts.

We would also like to answer the opposite question: "Can a team provide performance-related pay in a way that is both fair and economically viable?" However, the data in Lahman's Baseball Dataset are presented as summary statistics, and not play-by-play data. Should we acquire play-by-play or game-by-game data, they could be used to generate further insight into what factors make a team win. From there, we would be able to associate detailed team metrics with individual players to determine what makes a good baseball player and determine salary from those results. In lieu of play-by-play statistics, then, we answer a proxy question: "What in-game performance metrics are good predictors of wins?"

3 Data and Methods

We draw our insights using the raw data from Lahman's Baseball Database. Our relevant variables for our player salary analysis are: at bats (AB), hits (H), doubles (2B), triples (3B), home runs (HR), runs batted in (RBI), stolen bases (SB), caught stealing (CS), walks (BB), strike outs (SO), intentional walks (IBB), hit by pitch (HBP), sacrifice hits (SH), sacrifice flies (SF), and salary (response). The relevant variables for our analysis on game wins are the following: Team wins(W), team losses(L), earned runs(ER), runs scored(R), and fracwin(fraction of wins: $\frac{Wins}{Wins+Losses}$)

3.1 Principles of Measurement

In beginning our analysis, we identify principles of measurement that may affect the validity of the data or the subsequent analysis of data.

3.1.1 Distortion

We anticipate little distortion in the dataset, as baseball summary statistics are tabulated after plays have occurred; data collection would thus not affect results.

3.1.2 Relevance

Because Lahman's Baseball Dataset is so large, we anticipated several issues before we could begin exploratory analysis. Firstly, not all metrics are recorded for all of the 100+ years since data collection began. As such, there are several NaNs. To mitigate this issue and to make data easier to visualize, we opted to narrow the range of data analyzed to those taken in the 1990s. This is also when the current iteration of baseball (the designated hitter was introduced in the American League in the 1970s) was established enough to produce relevant data and insights to today's game.

After removing all data before 1990, we select the files in which data are relevant to the questions posed, namely "Salaries.csv," "Pitching.csv," "Batting.csv," and "FieldingOF.csv." Within these CSV files, we extract the relevant variables from their respective columns of data.

3.1.3 Precision

Most of the in-game performance data taken are for concrete, indisputable metrics (e.g., runs, hits). As such, they are only prone to data entry errors. However, other metrics, such as errors, are subject to human interpretation and must be taken with a grain of salt when producing insight.

3.1.4 Cost

Lahman's Baseball Database is free to use, but there are always costs associated with data collection. For every baseball team, a group must be hired and paid enough to motivate accurate data collection. Socially, there are other costs associated with data collection; it provides immutable evidence of how well players perform and as such, holds weight in the future career of all players. Care, then, should be taken to standardize data collection across the MLB.

3.2 Exploratory Analysis

In our exploratory analysis, we used many different diagnostic plots.

- Pair Plots: We used the pair plot from the seaborn library to analyze the distributions of the predictors and response, as well as observe the marginal relationships between variables.
- Heat maps: Another important diagnostic plot was the seaborn heat map of the correlation matrix. Computing a correlation matrix for all of the variables and plotting it on a heat map with a diverging color scheme allowed us to view negative and positive correlations between variables. Annotations over each cell demonstrated the numerical value for each correlation. This was a valuable resource because we were using regression to conduct our analyses.
- Scatter plots: Scatter plots are a useful tool in simple linear regression. In our analysis of team wins, the scatter plots and least squares regression line easily show the relationship between the response and predictor.

The pair plot was very useful in our analysis because it allowed us to see that many of our batting statistics had a vertical line of salaries at the 0 value for many of the batting statistics. After careful observation, it was apparent this line was representative of our pitching population in the batting dataset. This was reasonable because many pitchers are evaluated on their pitching abilities instead of batting statistics. After the removal of these data points, we halved the size of our sample. This mitigated visualization issues due to overplotting, and allowed us to view the players evaluated on their batting skills.

3.3 Predictive/Inferential Methods

In our analysis of the Lahman baseball data set, we used the statsmodel package, specifically the functions used for ordinary least squares regression. This included the use of the OLS function, which we used to create multiple models by least squares regression. From here, we used the predict method on our test data and observed the results of our model. The linear model would allow us to explore the relationship between variables in the dataset.

Additionally, we used many functions from the scikit learn library. For both of our regressions, we train-test split our data (75% Train, 25% Test). This allowed us to train our linear models and then use the testing data to attempt predictions. While overarching trends are reproducible using train-test split, one caveat is that the numbers will not be the same between runs as the function selects elements of the two sets randomly. For our principal component analysis, we used the PCA function. This function uses singular value decomposition to reduce the dimensions of the data and return (in our case, 10) principal components. From here, we were able to create a 2D plot of the first principal components and analyze differences in teams with different win percentages.

3.4 Predicting Player Salaries

To predict salaries based off of in-game performance, we perform pair plot analysis to determine relationships between explanatory variables. We then perform PCA to reduce the number of explanatory variables. Using the OLS function, we create a least-squares regression that represents the relationship between our predictors and the response, salary.

3.5 Predicting Wins from In-Game Performance

To predict wins from in-game performance, we look at the relationship between a team's statistics and their fraction of wins. Using the linear model we created from the data, we created predictions that would help us analyze which in-game performance metrics were good predictors of winning. Additionally, by using PCA, we could try to assess differences between teams that win more of their games and teams that lose more of their games.

4 Analysis

4.1 Assumptions

4.1.1 Salary Analysis

In our salary analysis, we looked at the data from the 1990 regular season. We chose this season because it was a nearly complete data set for the salaries of each player. Another reason that we chose the 1990 season's data is because choosing more than one year caused many issues with over plotting. Restricting our salary dataset to one year means we assume that fitting a model to the 1990 dataset will be representative of subsequent season and will be appropriate for salary predictions.

4.1.2 Team Wins Analysis

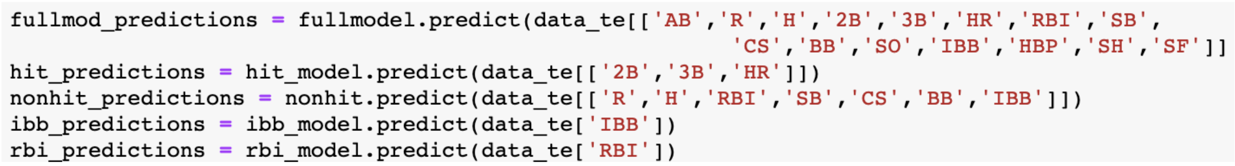
In our analysis of team wins, we looked at the statistics of baseball teams from 1990 to 2000. Choosing a subset of the entire population prevented over plotting, an issue we frequently encountered in our salary analysis. Using teams from 1990 to 2000 gave us a sample size of 278, which was enough to work with and prevent any extra noise from hiding relationships.

In this analysis, we were not given the statistics for individual games over the season. To make our analysis useful in a game setting, we assumed that the fraction of wins over the entire season was representative of the probability of winning a single game. This allowed us to use our analyses over entire season statistics to a game-by-game setting.

4.2 Analysis and Interpretation

4.2.1 Player Salaries

In our analysis of multiple batting statistics, we discovered that many of the models (**Figure 1**) were similar in terms of prediction accuracy. At a tolerance level of $100,000, all of our models predicted the test data values correct around 30 percent of the time. While only a small amount, the most ”accurate” model we explored was the intentional ball model. We hypothesized if a player was intentionally walked, then it was because the player is very skilled, and the pitcher does not want the batter to hit more than a single. In the analysis of our models, we discovered the biggest noticeable difference in the models was the percentage of over estimates and underestimates of the true values. While the SLR model of intentional balls was the most ”accurate,” it had the highest number of under estimates. This was most likely because it only took into account one single statistic as opposed to our full model. The RBI model, the hit model, and the non-full model were the next highest in underestimates, with the full model being the lowest.



```
fullmod_predictions = fullmodel.predict(data_te[['AB','R','H','2B','3B','HR','RBI','SB',
                                                       'CS','BB','SO','IBB','HBP','SH','SF']]
hit_predictions = hit_model.predict(data_te[['2B','3B','HR']])
nonhit_predictions = nonhit.predict(data_te[['R','H','RBI','SB','CS','BB','IBB']])
ibb_predictions = ibb_model.predict(data_te['IBB'])
rbi_predictions = rbi_model.predict(data_te['RBI'])
```

Figure 1: The linear models used in salary analysis used data from some columns of the raw datasets.

The number of underestimates could be useful in model selection. Teams would like to fit as many star players onto a team as possible, and using a model that underestimates is often valuable as the risk in associated with offering a player more money than offering them less is greater. If you offer a player more and they are a bust, then there is no space left to sign on another player for a contract. If the player is offered less and does not like the offer, there could be negotiating and agreement on a salary, or the player could take the initial offer and leave more monetary sources available for the team to use. One issue with our regression is we looked at one aspect of a player instead of looking at defense, leadership, and morale variables that can contribute to a players worth. Overall, the model has some legitimacy, but would need more predictors, or more analysis into useful predictors/model selection to give a better predictive model of player salary.

When we ran a regression of the model with doubles, triples, and home runs, all coefficients were were positive. This meant that for each coefficient, if the other coefficients were held constant, then a one unit increase in the observed coefficient would cause an increase in salary. This is intuitively correct. Against this intuition, any model that had ”hits” included made doubles, triples, and home runs negative. Whether hits was the only variable included or whether hits and other variables were included, all three coefficients ended up being negative. We ventured into interaction terms with hits and these predictors to see if there was any effect on the coefficients, but none was discovered. This was unprecedented in our analysis, and we were unable to find a way to fix this issue, or determine the underlying cause.

4.2.2 Team Wins

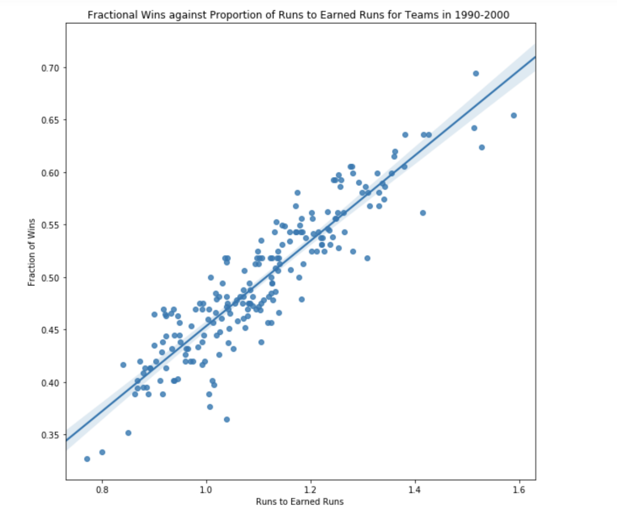


Figure 2: A regression line is fitted to fractional wins against proportion of runs to earned runs.

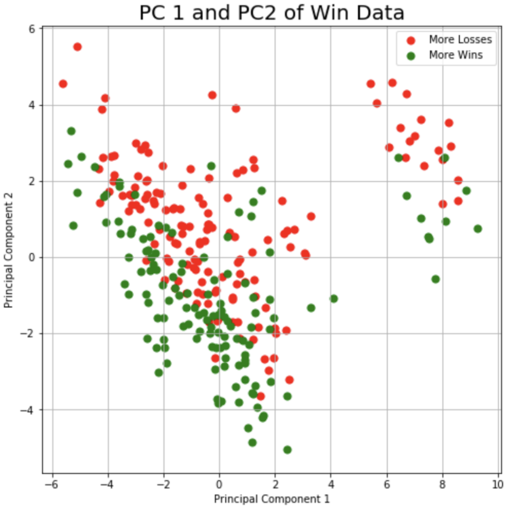


Figure 3: A principle component analysis shows two trends, one associated with winning teams and one associated with losing teams.

At a 5 percent tolerance, the model predicts the correct values about 90 percent of the time. At a 1 percent tolerance level, the model predicts the correct value about 30 percent of the time. This means that if we want a probability for a team winning a game within a margin of 5 percent based on the ratio of runs to earned runs, then our accuracy would be approximately 90 percent. If we wanted to predict the probability of a team winning within a margin of 1 percent, then we would be correct about 30 percent of the time. Under our assumption that this can be applied to a game scenario, we could predict the probability that a team is going to win based on the ratio of runs to earned runs. Our analysis into team wins was meant to give a predictive model that would allow in-game analysis into stats to predict the probability of a team winning. For this analysis, we decided to look at the ratio of runs to earned runs. For example, a value of 1.2 would mean that you are out scoring your opponents by 20 percent. According to our model, this corresponds to about a 51 percent probability of winning. With a model like this, teams could find the probability of winning a game based on how much they are outscoring their opponents. If the team has a comfortable margin then the coaching staff can switch out higher level players to give them rest so they are not overworked.

Observing the graph of the two principal components, there is a clear distinguishing of teams with more wins and teams with more losses on the line $y = -x - 2$. On the bottom portion of the line, there seems to be teams with more wins. Above the line there are more teams that have more losses. There is another grouping of points in the upper right corner from the bigger grouping of points. The extremely interesting part about this is that the groupings of points are in the same layout at the bigger grouping. The line $y = -\frac{5}{6}x + \frac{25}{3}$ divides the teams with more wins below the line, while the teams above the line are mostly teams with more losses. After some investigation, there were four expansion teams in the 1990's: Colorado Rockies(1993), Florida Marlins (1993) (Now Miami Marlins), Arizona Diamondbacks (1998), and Tampa Bay Devil Rays (1998) (Now Tampa Bay Rays). It is possible that because the analyzed data was from 1990 to 2000, this grouping represents these teams. Unfortunately, it is impossible to definitely say based on this visualization.

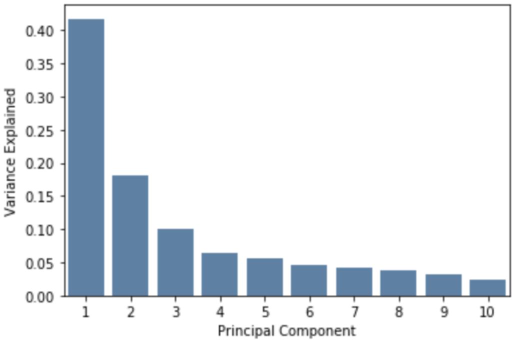


Figure 4: Scree Plot of the Win Data PCA

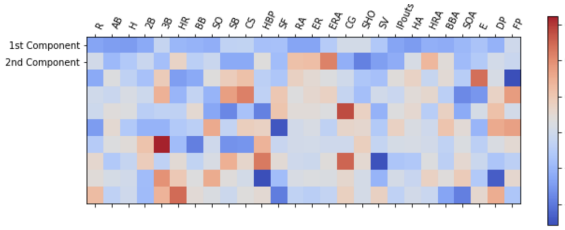


Figure 5: Composition of the All Components (First two are labeled)

The scree plot in **Figure 4** shows the percent of explained variance for each principle component. The first component explains about 42 percent of the variation, while the second component explains about 18 percent of the variation. The heat map shown in **Figure 5** shows how much of each variable accounted for each principal component. Noticeably, the second component seems to have been based highly off or Earned Run Average (ERA), Run Average (RA), Earned Runs (ER), Home Runs (HR), and Home Runs Allowed (HRA).

5 Performance-Related Pay – An Aside

In performing data analysis of baseball statistics and using insights to determine salaries, we are, by such actions, subscribing to performance-related pay (PRP). As its name suggests, performance-related pay rewards good performance with pay raises while minimally rewarding poor performance. There are both advocates and opponents of such a scheme in the business theory literature, but we side with the advocates, presenting both arguments below.

Advocates tout that PRP provides a means of standardization for the compensation of workers (or in the case of baseball, players). Performance expectations from executives are made clear, and favoritism in the form of salary boosts is eliminated. Using a PRP scheme, compensation is also robust to supervisor attitudes or increasing average productivity of a group of workers.

The largest criticism of PRP is that it reduces the myriad of tasks and duties associated with an occupation to a few measures of performance. Additionally, under PRP, little regard is given to the quality of tasks being performed. If PRP were used to incentivize workers to produce widgets faster, for example, it would foster a culture in which production and competition are valued over safety. Should accidents or sabotage between workers occur, it would undermine the long-term productivity of the company.

While we recognize the concerns of PRP opponents, we find that they do not apply well to the role of a professional baseball player. In particular, we see that the quality of a player's pitching, fielding, and batting can be assessed by concrete metrics; there is little room for subjectivity. Even for those metrics that appear gray, such as RBIs (i.e., a player may choose to bat differently based on how many players are on base and the particular bases on which they stand), there are few enough cases such that analysts can determine optimal hitting strategy and judge a player by level of execution.

6 Conclusion

We have demonstrated that data science can be used to drive useful insight for baseball team executives. In particular we show that three principal components are can explain a large majority (>70%) of the variation in salary. While we cannot use the metrics shown in Lahman's "Team.csv" file to produce a fair salary model based on a performance-related pay scheme, we demonstrate that there are non-obvious performance metrics that are good predictors of a team winning a particular game. Should Lahman's Baseball Dataset be augmented with play-by-play or game-by-game data, more insight can be used to propose an initial offering to free agents that is both competitive and in congruence with the player's performance.

References

[1] Maury Brown. Mlb spent less on player salaries despite record revenues in 2018. https://www.forbes.com/sites/maurybrown/2019/01/11/economic-data-shows-mlb-spent-less-on-player-salaries-compared-to-revenues-in-2018/#541b5a0339d7, 2018.