THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF ECONOMICS

INEFFICIENCIES IN THE MLB LABOR MARKET RELATED TO SALARY ARBITRATION

JOHN RUTHERFORD
SPRING 2019

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees
in Economics and Statistics
with honors in Economics

Reviewed and approved* by the following:

Dave Brown
Teaching Professor of Economics
Thesis Supervisor

Russell Chuderewicz
Teaching Professor of Economics
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

Professional sports provide opportunities to examine uniquely transparent labor markets. In particular, Major League Baseball exhibits a peculiar process known as salary arbitration. The Major League Baseball Collective Bargaining Agreement outlines this arbitration process, which is available to certain players who have not yet reached eligibility for free agency. This process elicits offers from both the team and the player for a one-year contract, and the arbitration panel assigns whichever offer they feel is closer to the true value of the player. This paper seeks to demonstrate the inherent opportunity for inefficiency that exists as a result of this arbitration process and quantify the consequential individual and league-wide misallocations of funds. As a means of analysis of these misallocations, salary will be modeled as a function of performance statistics for players eligible for free agency. These models will then be used to predict salaries for players that went through arbitration and players eligible for arbitration who avoided the process. These salary projections will be used to estimate the misallocations for players who go through arbitration and compare them to similar estimates for those players who avoided the arbitration process, thus providing a clear picture of the inefficiency of the arbitration process.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

To Dave Brown and James Tybout:

I sincerely thank you for your help throughout the process of developing and writing my honors thesis. Your advice and feedback have been crucial to the formation of this paper.

To my friends and family:

Thank you so much for always encouraging me to strive further in my academics to reach my potential. I truly would not have accomplished this feat without your constant encouragement and unwavering faith in my abilities. I will never be able to sufficiently describe how essential you have been to my success inside and outside of the classroom.

Chapter 1

Introduction

Summary of MLB Arbitration Process

Certain eligible players in Major League Baseball (MLB) may elect to enter a salary arbitration process when unable to reach an agreement for a contract with their team. Through this process, the player seeks a third-party decision for their compensation for the upcoming season. Both the team and player submit offers for a one-year contract, and the arbitration panel decides which offer is closer to what they believe to be the appropriate compensation for the player's services for that season. Based on the arbitration panel's decision, the closer offer is then imposed upon both parties.

Summary of Potential Inefficiencies of this Process

The arbitration process in the MLB as described above does allow for potential misallocation of funds. When neither party's offer is equal to the "true value" (as decided by the arbitration panel) of the player's services, the player will be inherently over- or under-paid, depending on which offer is closer to the true value. On a case-by-case basis, this difference may be small, but aggregated, there is the potential for large misallocations of funds within the MLB. This analysis will attempt to understand and estimate the individual and aggregate misallocations for players that went through arbitration prior to the 2018 MLB season.

Summary of Plan to Analyze for Possible Inefficiencies

In order to assess the extent of misallocation of funds through the arbitration process, I will first model salary as a function of performance statistics. Fielders and designated hitters will be modeled separately from pitchers due to the vast difference in the statistics for which these two types of players are evaluated. These models will be built using the 2018 salaries and 2017 performance statistics of all players who are eligible for free agency under the 2017 MLB Collective Bargaining Agreement (CBA). Only these players will be included in the model building process because they are the only players whose salaries are negotiated in a competitive market. The models will then be used to predict the "true value" salary for the players who have gone through the arbitration process. I will subsequently determine the difference between the predicted salary and the actual salary for each of these players. This will show the misallocations from the arbitration process. I will also evaluate the total misallocation for all players in order to gain a league-wide perspective of the inefficiency. Finally, I will follow the same procedure for the arbitration-eligible players who did not go through the arbitration process and compare their results to those who did go through arbitration. This will determine whether the possible inefficiency is result of arbitration or simply a problem in the evaluations of arbitration-eligible players.

Chapter 2

Literature Review and Negotiation Structure

Literature Review

In recent years, there has been an increase in interest in using data analytics to inform decision-making in sports. Due to the availability of various performance measures and large sample sizes, baseball has been one of the sports at the forefront of this data-driven revolution. Research has been conducted to answer various important questions regarding the determination of wages in Major League Baseball. Much of this research has specifically included examining the impact of components of the collective bargaining agreements that have been signed between the MLB teams and the players' union upon wages and team revenues. The relationship between player wages and the details of the CBA's exists because these agreements lay out the terms for player compensation structure.

Shifts in Importance of Statistics

In 2003, Michael Lewis published *Moneyball: The Art of Winning and Unfair Game*. In this book, he detailed how the Oakland Athletics had exploited a league-wide inefficiency in evaluating players to save money while maintaining success on the field. The other teams in the league were not placing enough weight upon a player's on-base percentage when evaluating and compensating players.

Brown, Link, and Rubin (2017) examine this inefficient environment and assert that in a competitive marketplace such as the MLB free agent market, other teams would change their behavior upon gaining knowledge that they had not been sufficiently accounting for on-base percentage when compensating players. Their analysis shows that prior to 2003, on-base percentage was not consistently a statistically significant factor in predicting the salary of a player who had gone through free agency, but since 2003, it has far more consistently been a statistically significant predictor of salary for free agent players. In their analysis, they also demonstrate that there is a significant interaction between on-base percentage and a binary variable indicating whether it is before or after the publication of *Moneyball* for free agent players. The significance and positive sign of this interaction shows that the importance of on-base percentage in determining player compensation has increased since the publication of *Moneyball*.

These analyses and conclusions from their article provide useful insights into which performance statistics are valued by MLB front offices in negotiating salaries in a free market. The authors also examine the possibility of temporal shifts in the importance of statistics and find significant evidence of these shifts. These changes over time are reasonable, as one would expect that as new information disseminates and new advanced performance measures are created, teams would take note and adjust their player evaluation methods accordingly.

Comparing Importance Between Statistics

Deli (2012) seeks to demonstrate issues in the processes utilized in some previous papers that have attempted to compare the importance of on-base percentage and slugging percentage. His analysis shows that neither simply comparing the regression coefficients of the two performance measures nor comparing their elasticities is sufficient to show that one measure is more important in determining a team’s offensive success (which he measures as runs per game).

Deli demonstrates that on-base percentage and slugging percentage cannot be compared in absolute terms because the statistics are measured in fundamentally different ways. On-base percentage theoretically ranges between 0 and 1 while slugging percentage theoretically ranges from 0 to 4. In practice, these statistics are concentrated toward the lower ends of their theoretical ranges. Drawing conclusions about importance by comparing regression coefficients would not yield accurate results because the statistics are measure on different scales. As Deli points out, a 0.01 increase in on-base percentage is a more substantial increase than a 0.01 increase in slugging percentage. He goes on to show that it is also improper to compare the importance of these statistics by examining the elasticities. A 1% increase in on-base percentage and a 1% increase in slugging percentage produce different amounts of change in the respective statistics because the variability of the statistics is not the same. Thus, Deli shows that some previously utilized methods of comparison do not truly produce the intended interpretations.

Deli then demonstrates that, when controlling for the difference in distribution between these two statistics, on-base percentage is not "more important." One important clarification that he makes is that his results do not show that the book *Moneyball* reached inaccurate conclusions about the Oakland Athletics' evaluation of on-base percentage. The main idea of that book is simply that on-base percentage was undervalued by the league for a time, not that it was necessarily more important than slugging percentage but was not treated as such.

The difference in the distributions of on-base percentage and slugging percentage is shown to complicate comparisons of deterministic importance between the two statistics. Deli shows that it is important to account for these differences in making any conclusions about which statistic is more important in determining runs scored or any other measure of success one would choose.

These results indicate the importance of caution when interpreting the regression results in this paper. It would be erroneous to assert that one statistic is more important than another based on a difference in coefficient estimate or elasticity because of the differing scales and distributions of these statistics. Thus, no conclusions about importance may be drawn from the magnitude or elasticity of a coefficient in these models.

Relationship Between CBA Changes and Player Salaries

As mentioned previously, the structure and contents of the collective bargaining agreements between players unions and sports leagues shape all interactions between players and team, including wage negotiations. Hill and Jolly (2017) address changes that occurred when the MLB and MLB Players Association (MLBPA) instituted a new collective bargaining agreement in 2007, and it establishes the effect that changes in CBA terms had on players' compensation.

Revenue sharing exists in baseball to ensure that all teams are interested in not only their own profitability but also the profitability of the other teams (Hill and Jolly 2017). It also ensures that teams that exist in smaller markets are granted a fair opportunity to succeed compared with the teams that operate in large markets.

Hill and Jolly's paper focuses specifically on changes in player salary brought about by the restructuring of the revenue sharing process that was laid out in the collective bargaining agreement signed in 2007. They explore the effects that the new CBA terms had on player salaries by dividing the players into two groups, position players and pitchers. Through their analysis, they find that both types of players experienced the theoretically expected increase in salary associated with the decrease in the revenue sharing "tax" brought about by the 2007 CBA. Interestingly though, they find that position players experience more of an increase than pitchers. On average, position players' salaries increased by 19.7% while pitchers' salaries increased by only 8.7%. They also found that there was no change in the weight given to performance statistics when setting salaries after the institution of the new CBA.

These results demonstrate the importance of the terms of the CBA in MLB salary negotiations. They show that a change in terms of the CBA does not necessarily indicate a change in the way teams value certain performance metrics; however, a change can still affect the compensation of players. Depending on the terms of the two collective bargaining agreements since the one signed in 2007, similar phenomena may have occurred again since the period examined.

These results show the importance of the influence of the CBA in determination of player compensation. While their focus was on the effect of CBA changes, specifically regarding revenue sharing, this analysis will focus on the compensation and arbitration process as detailed in the 2017 MLB CBA. The connection of this aspect of the CBA to player compensation is far more obvious because it defines the ability of players to negotiate their wages with a team or teams.

Effects of Third-Party Impositions on Long-Term Relationships in Baseball

The MLB labor market provides a suitable environment for examining the sustainability of relationships after either a voluntary agreement or third-party-imposed resolution are reached. Budd, Sojourner, and Jung (2017) set out to investigate whether relationships between players and teams are significantly impacted by the type of agreement reach: voluntary or imposed.

In the context of this study, a voluntary agreement is a contract mutually agreed upon by the team and player while an imposed resolution is an arbitration-imposed contract. It would seem that the imposed resolutions in this case would tend to cause more relationship breakdown because they are the product of the two parties being unable to reach an agreement for a contract.

This study substantiated the idea of arbitration-caused relationship breakdown through its analysis. They found that "players who experience arbitration…are significantly less likely to still be with the arbitration team at the end of the season immediately following arbitration, even after we control for player quality and the intensity of the disagreement." Specifically, the study found that, on average, players who went through the arbitration process were 7.3% more likely to be released or traded by the team that year if the player lost the hearing and 15.2% more likely to be released or traded if the player won.

Thus, while the arbitration process is structured to ensure fairness, the imposition of a third-party decision ultimately hurts the relationship between team and player. This would seem to be due to the adversarial nature of the negotiation and arbitration process and the fact that the arbitration process either pays players less than they feel they deserve or more than the team feels the player is worth. This first instance would likely cause unrest in a player who felt they were being insufficiently compensated, which could lead to a request for release or trade. The second possibility would incentivize the team to get the salary off of the books because the management feels that the player's production is not worth the price.

This study shows how the arbitration process can be regarded by both teams and players as an especially adversarial environment and lead to the breakdown of the relationship between teams and players. This gives support to the idea that MLB players and teams would like to avoid arbitration when possible to both protect against the risk of losing the hearing and to avoid a breakdown in the relationship between the two parties.

Risk Aversion and Long-Term Contracts in the MLB

Long-term contracts offer protection from risk for both players and teams. For players, the possibility of catastrophic or even career ending injury is ever-present. This could jeopardize the earnings that the player would anticipate if he was able to play for a longer period of time. For teams, changes in the market could lead to a team being forced to pay more than they would like in order to keep a player. Long-term contract extension can mitigate these risks for both parties because they provide income security for the player and protect against changes in costs for the team. This concept is similar to that of a futures contract, but rather than protecting against the change of a price in a good or commodity, the parties are protecting themselves from a change in the value of a player's services.

Walters, von Allmen, and Krautmann (2017) set out to analyze 106 player contracts to examine the possible risk-averse behavior exhibited by teams and players in the negotiation of long-term contracts. The authors find that the contracts exhibit more risk-averse behavior on the part of the teams than the players. They found that teams paid an average of about $571,000 per player per year more than the team would have paid for consecutive one-year agreements, behaving especially cautiously when negotiating with their "key players". This demonstrates that teams tended to take a cautious approach in order to protect against larger costs in the future. Players were also risk-averse in many cases, and "almost 30% appear to have signed agreements that gave up significant prospective income on year- to-year contracts in order to guarantee smaller returns."

These results show that both players and teams often engage in risk-averse behavior in order to protect themselves from the possibility of hurtful price fluctuations in the labor market. It is important to consider that my analysis will not treat long term contracts differently than short term contract. Compensation will simply be evaluated as a player's 2018 salary and the length of the contract will not be examined. Thus, it is important to assume that the risk-averse behavior will be proportional over the entire sample of player salaries to be examined, with no party taking drastically more risk-averse stances than the other.

Brief Overview of MLB Salary Negotiation and Possible Inefficiencies

Eligibility Classification Rules

Salary level in the MLB is based on a three-class system of payment. Players fall into one of three classifications: pre-arbitration players, arbitration players, and free agent players. Players' ability to negotiate the terms of their compensation are based upon their class (Collective Bargaining Agreement, 2016).

The length of time that a player has spent in the MLB determines which class that player belongs to. One "year of credited service" is defined as 172 days on a team's Active List (Collective Bargaining Agreement, 2016). The number of days from one season that may count toward a player's year of credited service total is capped at 172, so one player may accrue a maximum of exactly one year of service during one MLB season (Collective Bargaining Agreement, 2016). Players with less than three years of credited service are classified as pre-arbitration players. Those with at least three but less than six years of credited service are classified as arbitration players with one exception known as the "Super Two" being classified as an arbitration player after only two years of credited service. Finally, players with six or more years of credited service are classified as free agent players.

Pre-arbitration players are commonly referred to as "team-controlled" players because of their lack of power to negotiate their compensation. Teams are able to compensate pre-arbitration players as they see fit. Moreover, many simply employ a universal rule that determines all pre-arbitration players' salaries, sometimes based on years of credited service or performance statistics. Teams have this power because pre-arbitration players are not able to seek better compensation from other teams when their contracts expire. They are bound to the team that they were previously signed with until they attain eligibility for free agency.

Arbitration players experience circumstances similar to those described above for pre-arbitration players. The distinguishing property for arbitration players is their ability to enter in to the arbitration process to seek a better wage from their team. If an arbitration player and their team are unable to reach an agreement for a new contract, the player may choose to file for arbitration. This process is led by a panel which uses comparable players' compensations to determine a fitting salary for a one-year contract for the filing player. Both the players' representation and the team propose a one-year salary offer to the panel. Whichever is closer to the fitting salary that the panel found independently is imposed on the team and player. Thus, the arbitration process can yield a higher salary in favor of the player or a lower salary in favor of the team depending on the offers submitted by each party and the panel's independently assessed "true value" of the player. Due to this uncertainty, many arbitration players may file for arbitration to pressure their team to compromise but ultimately sign before their hearing to avoid the uncertainty of a third-party imposed salary.

Free agent players experience the most freedom in their salary negotiations. Upon the expiration of the contract of a free agent player, the player may negotiate and sign with any team. This allows for a competitive market of all MLB teams to determine the compensation plan deemed suitable for each player. Thus, free agent players are compensated most fairly for their past production and future projections. Because free agent players have had six credited years of service to demonstrate their abilities, teams have a better understanding of the on-field production that they can expect from each player, so there is not an asymmetry of information between the players and teams.

Potential for Inefficiencies

The classification rules outlined above demonstrate the potential for inefficiencies in the outlined MLB wage negotiation process. Pre-arbitration players can be compensated far less than their production warrants because of their powerlessness in determining their compensation. Arbitration players can experience that same inefficiency or experience a different problem as a result of the arbitration process. This potential misallocation due to the arbitration process will be the focus of this paper.

The first duty of the arbitration panel is to determine the salary that each filing player should be paid for a one-year contract. The panel bases this on performance and uses comparable players' salaries to link performance to payment. For this analysis, the arbitration panel's initial estimate of appropriate salary will be assumed to be the salary a player would receive in the free agent market.

The team and player's representation are then responsible for filing their offers for a one-year contract, which will ideally be as far from the panel's figure as possible while remaining closer than the offer of the opposing party. Each party is trying to maximize their distance from the "true value" as estimated by the arbitration panel subject to the constraint of being closer than the other party. This characteristic of the arbitration process has vast potential for misallocation of wages. If the team wins with an offer that is below the "true value," the player will be under-paid; however, if the player wins with an offer that is above the "true value," the player will be over-paid. Any result that does not end in one party offering the value proposed by the panel is inefficient. The aggregation of each of these potential inefficiencies could balance out, where players are over-paid about as often as they are under-paid. The more concerning possibility is that one outcome is far more frequent than the other. This would mean that the arbitration process is either costing teams money by over-paying players who do not deserve the extent of their compensation, or it is costing players money by under-valuing their performance and failing to compensate them fairly and accurately. In either scenario, it would be important for the MLB to consider restructuring their salary negotiation process. The apparent necessity of any steps by the MLB will be addressed by the analysis conducted in this paper.

Chapter 3

Methodology

Data Overview

Data Collection Methods

In order to perform the analysis plan summarized previously, data regarding players' performance statistics for one season and their salaries for the next season will be necessary. Data for players' 2018 salaries was gathered from a USA Today database. Various performance statistics for players' 2017 seasons were collected from USA Today and Baseball Reference, and arbitration eligibility data was collected from Spotrac and Baseball Prospectus. By merging information from these datasets, I created new datasets containing players' 2018 salary, 2018 salary negotiation classification, and 2017 performance statistics. The datasets are divided by player type: pitcher or hitter. Pitchers were not evaluated as hitters regardless of their league affiliation (National League pitchers hit while they do not in the American League), and hitters were not evaluated as pitchers. Thus, no player appears in both datasets. These datasets allow for model building using players eligible for free agency and prediction of arbitration players using this model.

Definition of Populations

There is a total of 416 fielders/designated hitters and 461 pitchers. The datasets contain only those players included in the USA Today salary dataset, as that was the more exclusive than the dataset containing player performance statistics. The USA today population is defined as all players listed on the 25-man roster or disabled list of all MLB teams on the first day of the season (i.e. Opening Day).

Variable Descriptions

For each player, various attributes are listed for identification and analysis purposes. Some attributes are used for both groups, but the majority are distinct based on the group to which the players belong due to evaluation differences. The characteristics recorded for all players are name, salary, eligibility classification (free agency eligible, arbitration avoided, arbitration settlement, or pre-arbitration), position, team, and games played. Specific variables for fielders/designated hitters and pitchers are listed and defined in the tables below.

Table 1. Hitting Performance Statistics and Descriptions

Statistic	*Description*
At-Bats (AB)	number of times a player has an opportunity to hit, not including walks or hit by pitches
Runs (R)	runs scored by a player
Hits (H)	total number of hits
Doubles (2B), Triples (3B), Home Runs (HR)	total number of each type of hit respectively
Runs Batted In (RBI)	number of runs scored due to the player's batting outcome
Total Bases (TB)	TB = Hits + Doubles + 2*Triples + 3*Home Runs
Walks (BB)	total number of times a player reached base on a walk by the pitcher
Strikeouts (SO or K)	total number of times a player is called out on strikes
Stolen Bases (SB)	total number of successfully stolen bases
Caught Stealing (CS)	total number of times that a player is unsuccessful in stealing a base
Batting Average (BA)	BA = Hits/At-Bats

On-Base Percentage (OBP)	OBP = (Hits +Walks + Hit by Pitches)/(At-Bats + Walks + Hit by Pitches)
Slugging Percentage (SLG)	SLG = Total Bases/At-Bats
On-Base Plus Slugging (OPS)	OPS = OBP + SLG

Table 2. Pitching Performance Statistics and Descriptions

Statistic	***Description***
Starts (GS)	total number of games where the player is the first pitcher to play for their team
Innings Pitched (IP)	total number of innings pitched by the player
Hits Allowed (H)	total number of hits rendered to opposing teams by the pitcher
Runs Allowed (R)	total number of runs rendered to opposing teams by a pitcher
Earned Runs Allowed (ER)	total number of runs rendered to opposing teams by a pitcher which are not the results of fielding errors
Walks Allowed (BB)	total number of walks rendered to opposing teams by a pitcher
Strikeouts (SO or K)	total number of strike outs earned against opposing teams by a pitcher

Strikeouts per Nine Innings (K/9)	K/9 = (Strikeouts/Innings Pitched)*9
Wins (W)	total number of games where the pitcher is deemed by rule to be the winning pitcher
Losses (L)	total number of games where the pitcher is deemed by rule to be the losing pitcher
Saves (SV)	total number of save opportunities where the pitcher earns a save
Blown Saves (BLSV)	total number of save opportunities where the pitcher does not earn a save
Walks and Hits per Inning Pitched (WHIP)	WHIP = (Walks + Hits)/Innings Pitched
Earned Runs Average (ERA)	ERA = (Earned Runs/Innings Pitched)*9

Modeling Overview

Modeling Approach

As outlined previously, in order to examine the discrepancy between the "true value" of players who have gone through arbitration and their actual compensation, I will construct models predicting 2018 salary using 2017 performance statistics. The models will be built using data for only players eligible for free agency, then they will be used to measure the misallocation for each player that went through arbitration. Additionally, the models will measure the misallocation for arbitration-eligible players who did not go through the arbitration process. This will allow observation of the differences in compensation and misallocations between those who do and do not go through arbitration. In order to construct these models, I will use Ordinary Least Squares (OLS) regression methods. Various models with different predictors for hitters and pitchers will be evaluated and included for calculation of inefficiencies.

Transformation of Salary

In order to conform to the assumptions of Ordinary Least Squares regression, the response of salary must be normally distributed. In order to see whether transformation of salary would be necessary, I first created Box-Cox Plots to see what transformation if any was suggested (Appendix A). Based on the intervals and suggested transformations in these plots, I deemed it appropriate to evaluate both untransformed salary and the natural logarithm of salary as possible predictors. Additionally, due to the presence of a salary minimum in the MLB, I decided to evaluate the feasibility of excluding players in the top 10% for salary to create a more normal distribution. Each of these possible responses would be considered for hitters and pitchers separately, but for ease of interpretability and drawing conclusions, the same transformation would be used for both types of models.

As a means to assess the normality of each of these possible responses, I created histograms and QQ-plots (Appendix B). Each of these graphs communicated important information about the normality of the variable, with histograms giving an image of the distribution and QQ-plots showing the closeness of the distribution to a normal distribution. Based on these plots, I found that the natural logarithm of salary would be the most suitable response for my modeling efforts. Thus, all models would be built using this transformation.

Evaluation of Relief Pitchers

Relief pitchers serve a different role than starting pitchers. Rather than being relied on heavily to pitch a large number of innings, they are typically brought in situationally to perform a more specific job. Thus, they may be evaluated differently for some statistics than starting pitchers are. Also, they are evaluated according to some additional statistics not used to evaluate starting pitchers. To account for these differences, an indicator variable for relief pitcher will be introduced, and the interaction of this variable and the pitcher performance statistics (e.g. saves or SV) will be evaluated in the model. This will account for the fact that many starting pitchers will have few or no saves and allow for analysis of the importance of these variables that are specific to relief pitchers, as well as allow analysis as to whether innings pitched is less important for relief pitchers than for starters.

Variable Selection Restrictions

Many of the variables in the dataset are related and/or used to calculate other variables in the dataset. For this reason, it is not sensible to include certain variables in the same model. For both hitters and pitchers, certain variables will not be included together to eliminate problems that would arise from these simultaneous inclusions.

For hitters, batting average (BA) is calculated as hits (H) divided by at-bats (AB), so these variables should not be included together. On-Base Percentage (OBP) includes all components of BA, thus it gives the same information as BA along with additional information specific to OBP. Walks (BB), hits, and at-bats are each components of OBP, so they will not be included with OBP. Each type of hit (double, triple, home run), the hits variable itself, and at-bats are components of slugging percentage (SLG), so these variables will not be included with SLG.

Pitchers statistics have similar limitations to hitters. Earned runs (ER) divide by innings pitched (IP) is used to calculate earned runs average (ERA), so both of these variables will not be included together with ERA. Similarly, hits (H), walks (BB), and innings pitched are used to calculate walks and hits per inning pitched (WHIP), so they will not all be included with WHIP. Finally, strikeouts per nine innings (Kper9) is calculated using strikeouts (SO) and innings pitched, so all three will not be simultaneously included.

Chapter 4

Empirical Results

Summarize Model Results

Table 3. Model Results for Hitters (coefficient estimate, p-value, VIF)

	Intercept	*GP*	*AB*	*R*	*H*	*RBI*	*BB*	*SO*	*AVG*	*OBP*	*SLG*
Model 1 (R-sq = 5.596%)	13.8426 <2e-16* -									5.7926 0.00556* 1	
Model 2 (R-sq = 14.68%)	14.2047 <2e-16* -									-1.4889 0.58979 1.973279	4.8430 0.00025* 1.973279
Model 3 (R-sq = 36.73%)	12.700046 <2e-16* -						0.015224 5.23e-6* 1.176458		6.441952 0.00038* 1.236489		1.829833 0.03922* 1.231238
Model 4 (R-sq = 46.38%)	13.0759824 <2e-16* -		0.0042445 7.94e-9* 2.471990					-0.0013577 0.635772 2.392038			2.6939674 0.000484* 1.066253
Model 5 (R-sq = 49.93%)	14.733404 <2e-16* -	-0.018402 0.000229* 7.358947		-0.017557 0.037570* 11.873693	0.026235 2.91e-6* 13.879179	0.006196 0.266371 5.804337	0.013082 0.011655* 3.695670	0.006082 0.067734* 3.335184			

Table 4. Model Results for Pitchers (coefficient estimate, p-value, VIF)

	Intercept	*GP*	*IP*	*Kper9*	*W*	*L*	*relief:SV*	*WHIP*	*ERA*	*relief:IP*
Model 1 (R-sq = 11.45%)	16.67143 <2e-16* -								-0.24506 0.000129* 1	
Model 2 (R-sq = 12.49%)	17.114564 <2e-16* -			0.007529 0.876 1.387437				-0.847701 0.261 4.703311	-0.104952 0.432 4.601534	
Model 3 (R-sq = 12.9%)	17.310939 <2e-16* -						-0.008143 0.447 1.073240	-0.894235 0.227 4.552571	-0.115967 0.383 4.577279	
Model 4 (R-sq = 45.43%)	17.321059 <2e-16* -		0.003855 0.0163* 1.879446					-0.795610 0.1831 4.685949	-0.0156289 0.1402 4.565779	-0.012550 3.87e-5* 1.903726
Model 5 (R-sq = 41.96%)	16.885189 <2e-16* -	-0.012206 0.01460* 1.835075			0.065002 0.00293* 2.152645	0.039884 0.07837* 1.800035	0.016741 0.08093* 1.248450	-0.498862 0.41773 4.617365	-0.184387 0.11353 5.094398	

The results of various OLS regression models built using data for players eligible for free agency are summarized in Tables 3 and 4 above. Asterisks next to the p-value of a variable indicate statistical significance of that variable at a 10% level of significance. Under each model is listed the R-Squared value of that model, which is the percentage of variation in the natural logarithm of salary that is explained by the selected predictor variables.

Due to the close relation of the statistics included in the models examined, correct interpretations of the coefficient estimates are important. In Model 1 for hitters, the coefficient for OBP is positive, indicating the intuitive conclusion that higher OBP results in a higher logarithm of wage. However, in Model 2, the coefficient estimate for OBP is negative. This is a result of the inclusion of SLG in the model. These results show that the logarithm of salary increases with SLG at a constant level of OBP, but that the logarithm of salary decreases as OBP increases at a constant level of SLG. Model 3 for hitters shows that walks, batting average, and slugging percentage are each significantly and positively related the response. Model 4 shows that at-bats and SLG are significant with a positive coefficient. These results are again very sensible, as higher at-bats would indicate that a player is playing more often and is deserving of a higher salary. Finally, for hitters, Model 5 shows that GP and R have significant, negative coefficients, but that H, BB, and SO have significant positive coefficients.

For pitchers, it is similarly important to correctly interpret the coefficients. High ERA indicates worse pitching performance, so the significant negative coefficient in Model 1 is unsurprising. Model 2 and Model 3 do not offer any insight about coefficients because none of the coefficients in these models are found to be significantly different from zero. Model 4 shows that innings pitched is significant and positively related to the response, which is reasonable because pitchers performing better would be left in games longer. Additionally, there is a significant interaction between the relief pitcher indicator variable and innings pitched, which shows that the effect of innings pitched is significant less for relievers than starters due to its significant negative coefficient. This result is again reasonable because relief pitchers are not expected to pitch as many innings as a starter would. Lastly, Model 5 shows that the coefficient on GP is significantly less than 0, while the coefficients for wins, losses, and the interaction between saves and the relief pitcher variable are each significant and positive.

These model results (specifically the coefficient estimates) will be used to examine the difference between arbitration-eligible players' compensation and their predicted compensation based on the models. This will show how these players are compensated relative to their worth on a free market.

Total Misallocation for Players That Went Through Arbitration

In order to calculate estimates of the league-wide misallocation of funds through the arbitration process, I first predicted the salary for each player that went through arbitration using the models discussed previously. Then, to find the individual misallocation, I subtracted the predicted salary from the player's true salary, which yields the misallocation for that individual. The sum of each individual misallocation gives and estimate for the league-wide misallocation. This estimate was calculated for each pair of hitter and pitcher models. These estimates are displayed in Table 5 below. This table shows that all estimates for the league-wide misallocations show that the players are being under-compensated relative to their performance when they go through the arbitration process. This does not, however, show that the arbitration process is the cause of the under-compensation. The minimum estimate for the total misallocation across the MLB is -30,769,452 while the maximum estimate is -103,856,294. Thus, the estimates for the league-wide misallocation range from players being under-compensated by a total of $30,769,452 to players being undercompensated by a total of $103,856,294. This is especially jarring in light of the fact that only 19 players went through arbitration for the 2018 season, so on average, these players have an estimated under-compensation between $1,619,444.84 and $5,466,120.74 each.

Table 5. Estimates of Total Misallocation for Players that Went Through Arbitration

	Hitter Model 1	Hitter Model 2	Hitter Model 3	Hitter Model 4	Hitter Model 5
Pitcher Model 1	-36,422,319	-49,841,133	-50,667,040	-71,292,175	-74,441,009
Pitcher Model 2	-34,569,288	-47,988,102	-48,814,009	-69,439,144	-72,587,979
Pitcher Model 3	-30,769,288	-44,188,266	-45,014,173	-65,639,308	-68,788,143
Pitcher Model 4	-58,294,500	-71,713,314	-72,539,222	-93,164,357	-96,313,191
Pitcher Model 5	-65,837,603	-79,256,417	-80,082,325	-100,707,460	-103,856,294

Average Misallocation for Players that Avoided Arbitration

In order to best interpret the misallocations for the players that went through arbitration, it is useful to provide context by also examining those players who have avoided the arbitration process to see whether they have achieved different results. This makes it clearer what effect, if any, the arbitration process has on the under-compensation of players that has been estimated. In order to compare the players that went through arbitration to those who avoided arbitration, it is easiest to calculate the average misallocation estimate for each model using data for players who avoided arbitration. Comparing these estimates to the average misallocation for players who went through arbitration will show whether players are receiving more or less favorable results via the arbitration process. Table 6 below summarizes the estimate for average misallocation for players who avoided arbitration and those who went through arbitration.

Table 6. Average Misallocation for Players who Avoided or Went Through Arbitration

	Hitter Model 1	Hitter Model 2	Hitter Model 3	Hitter Model 4	Hitter Model 5
Pitcher Model 1	-2,653,747 -1,916,964	-2,920,495 -2,623,218	-3,059,952 -2,666,686	-3,814,450 -3,752,220	-3,980,770 -3,917,948
Pitcher Model 2	-2,509,423 -1,819,436	-2,776,172 -2,525,690	-2,915,628 -2,569,158	-3,670,126 -3,654,692	-3,836,446 -3,820,420
Pitcher Model 3	-2,491,811 -1,619,445	-2,758,560 -2,325,698	-2,898,016 -2,369,167	-3,652,514 -3,454,700	-3,818,834 -3,620,429
Pitcher Model 4	-2,552,737 -3,068,132	-2,819,486 -3,774,385	-2,958,942 -3,817,854	-3,713,440 -4,903,387	-3,879,760 -5,069,115
Pitcher Model 5	-2,509,976 -3,465,137	-2,776,725 -4,171,390	-2,916,181 -4,214,859	-3,670,679 -5,300,393	-3,836,999 -5,466,121

In Table 6 above, the first number in each entry is the average misallocation for a player who avoided arbitration, and the second number is the average misallocation for a player who went through arbitration. Upon examination of these estimates, we see that players who avoided arbitration are similarly under-compensated to those who went through arbitration. This would indicate that the arbitration process is not the cause of this under-compensation, but that it is instead cause by league-wide under-valuing of players before they reach free agency.

Chapter 5

Conclusion

The goal of this analysis was to examine the inefficiencies in the salary negotiation structure of the modern MLB, specifically those inefficiencies related to the arbitration process. In order to determine the extent of the inefficiency, models were constructed to predict a player's salary based on their performance in the prior season. These models were then used to determine the relationship between arbitration-eligible players' salaries and their predicted salaries. The difference between these two values would be the estimate for the inefficiency or misallocation for that player. These misallocations were then aggregated and average to give a better picture of the league-wide trends related to arbitration.

Summary of Arbitration-Related Inefficiencies

This analysis yielded several meaningful conclusions about both the implications of the arbitration process and the state of MLB salary negotiations for arbitration-eligible players in general. It first showed how, on a case by case basis and league-wide, arbitration yields inefficient outcomes for both parties involved. Additionally, it showed that players who avoid arbitration are similarly under-compensated for their play. This indicates that the MLB pays arbitration-eligible players less than they would if they were free agents. This is an unsurprising result because the salary negotiation structure is constructed to compensate players more fairly upon reaching free agency.

Based on the results for every model examined, players are, on average, underpaid through the arbitration process based on their production from the previous season. The estimates of the sum of these misallocations ranged from $30,769,452 to $103,856,294 for the 2018 season. Considering that there were only 19 players that went through the arbitration process for this season, it is alarming that the misallocations would be so high. It is also noteworthy that for every model examined, the total misallocation was negative, indicating that the teams were the party that benefits from arbitration process and the salary negotiation structure.

When evaluating the average misallocation for players that avoided arbitration, there were similar results to those who had gone through the arbitration process. The mean misallocation for each pair of models was similar for each set, and the dissimilarities may be due to the small sample size of only 19 players going through the arbitration process. These similarities in under-compensation between players who did and did not go through arbitration indicate that the arbitration process is likely not the main cause of the misallocations for arbitration-eligible players. Thus, the team-friendly nature of the salary negotiation process seems to extend beyond the arbitration process to encompass all arbitration-eligible players. This shows that the league as a whole underpays these players on average.

Limitations and Considerations

Due to the methods used to analyze this data and that dataset itself, there are limitations to the results of this analysis. There is missing data in this dataset because 46 pitchers and 12 hitters that were listed in the salary dataset did not have statistics for the 2017 season due to injury or other extenuating circumstances. Thus, not every player in the defined population was classified and modeled. Also, due to the extensive nature of the data collection process, only the prior year of performance statistics was examined. Examining beyond one year prior becomes more difficult because fewer current players will have performance statistics for a year as one gets data from further in the past. Another source of possible omitted variable bias is the fact that no advanced statistics were used in the modeling process (e.g. WAR, FIP, BABIP). The lack of adjustment for long-term contracts versus short-term contracts is also a factor that could affect the results of this analysis.

Limitations with the modeling procedure specifically are associated with the normality of the predictor and multicollinearity of the variables. Salary did not appear to be a suitable response for OLS because it did not appear approximately normal. To account for this, the natural logarithm was used instead. This transformation was not perfectly normal either, but appeared to be an improvement and adequately normal. Some of the predictor variables also, as outlined previously, do have inherent relationships and correlations, but based on the variance inflation factors generated in the results of the code, the multicollinearity did not appear to have a strong effect on the results. The Variable Selection Restrictions sought to limit the negative impact of these predictor relationships, and it appears as if these effects were well-mitigated.

Future Steps

Future analysis building off of this paper could provide a more insightful look into the inefficiency of MLB salary negotiation. A major step for future analysis would be to include more advanced performance statistics in the models (e.g. wins above replacement, fielder independent pitching). It would also be helpful to include other additional variables such as age and performance statistics for seasons earlier the previous season. Additionally, including multiple years of salary and performance data while accounting for inflation and other economic conditions affecting the sport of baseball could give a model that could be helpful in the long-term as a tool for evaluating the worth of player performance. Differentiating between short-term and long-term contracts would also provide more information about how teams value players. Based on the structure of the arbitration process as it is outlined currently, it could also be modeled as a game of asymmetric information. The equilibria of such a game would prove useful to both parties involved in the arbitration process. Proposals of new structures for salary negotiation before free agency could also give the MLB an opportunity to evaluate new possibilities. With threats of a labor strike looming, an alternative to this seemingly inefficient process could bring the parties of the league and the players union closer to an agreement that could avoid such a strike.

Appendix A

Box-Cox Plots of Salary for Hitters and Pitchers

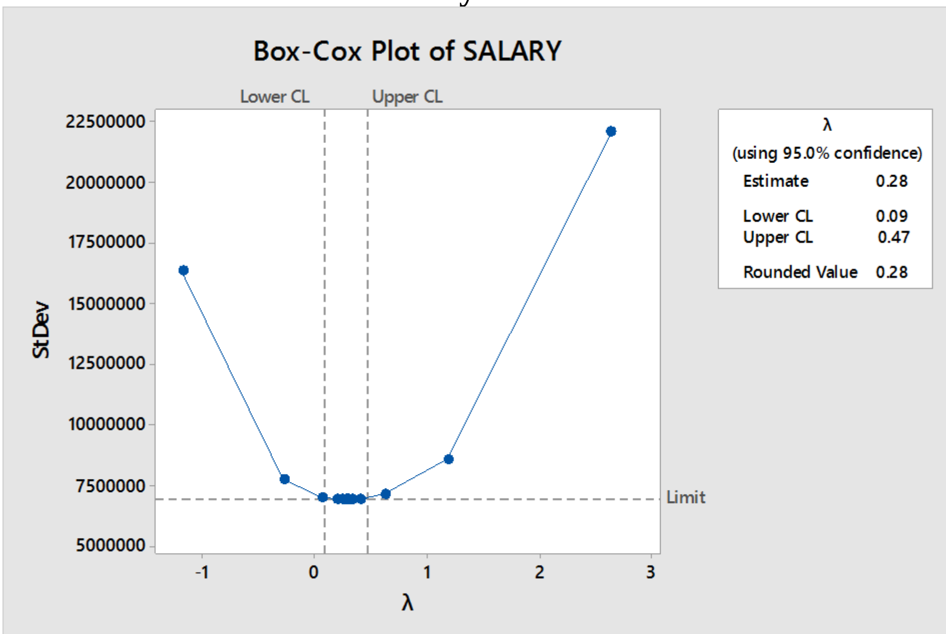

Figure 1. Box-Cox Plot for Hitters Eligible for Free Agency

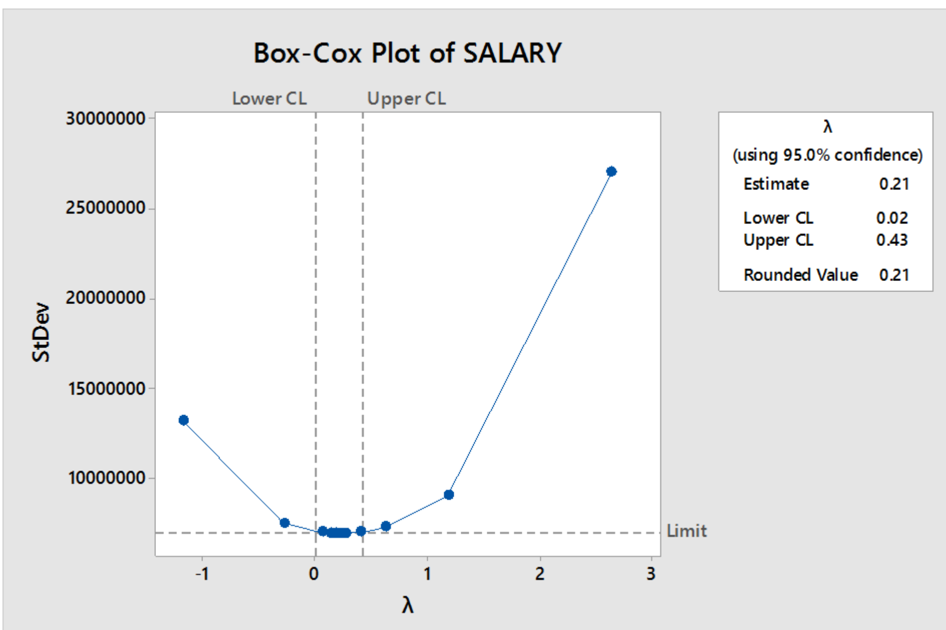

Figure 2. Box-Cox Plot for Pitchers Eligible for Free Agency

Appendix B

Histograms and QQ-Plots for Salary and Transformations of Salary

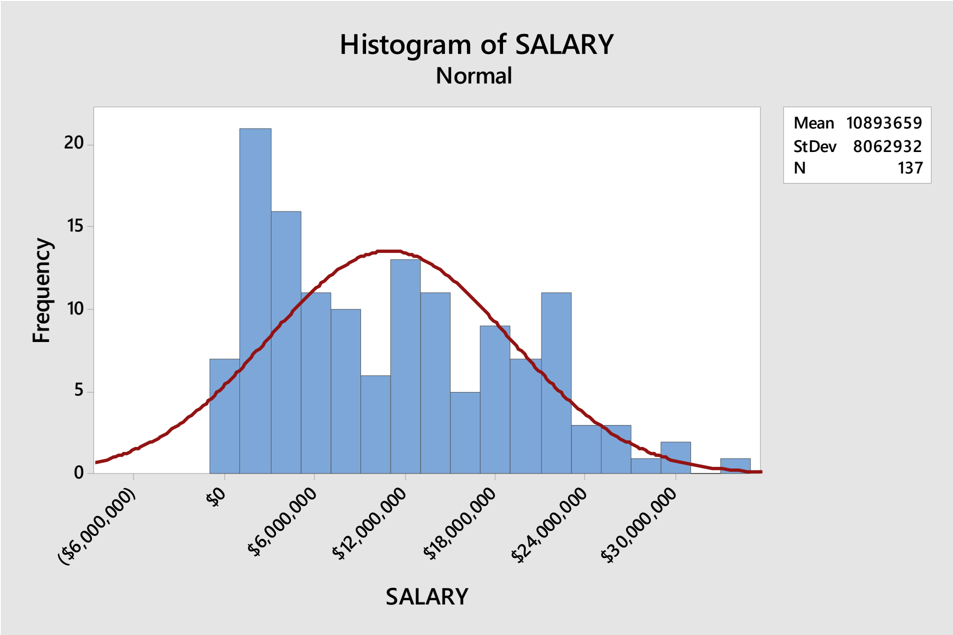

Figure 3. Histogram of Salary for Hitters Eligible for Free Agency

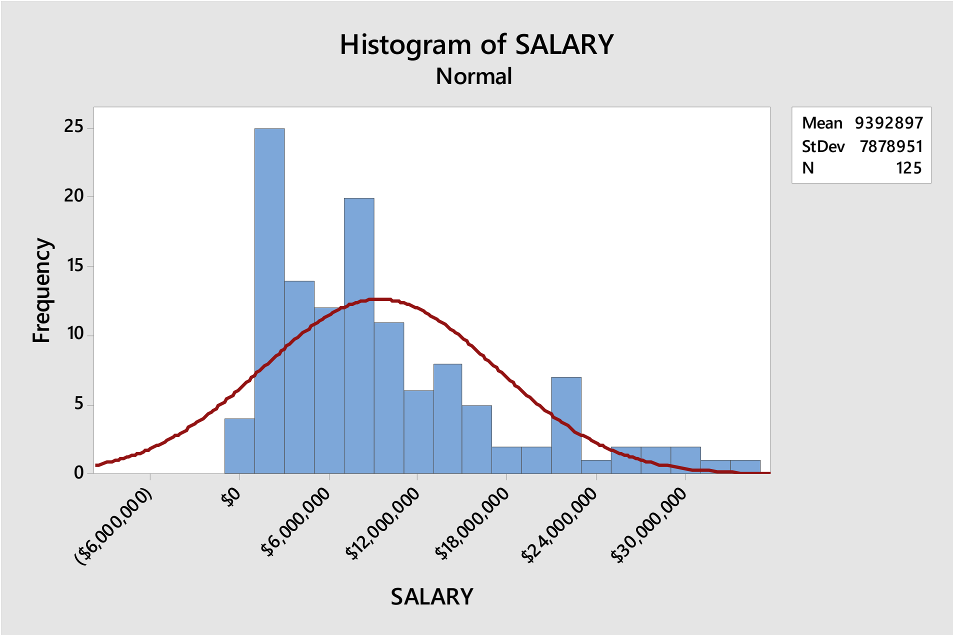

Figure 4. Histogram of Salary for Pitchers Eligible for Free Agency

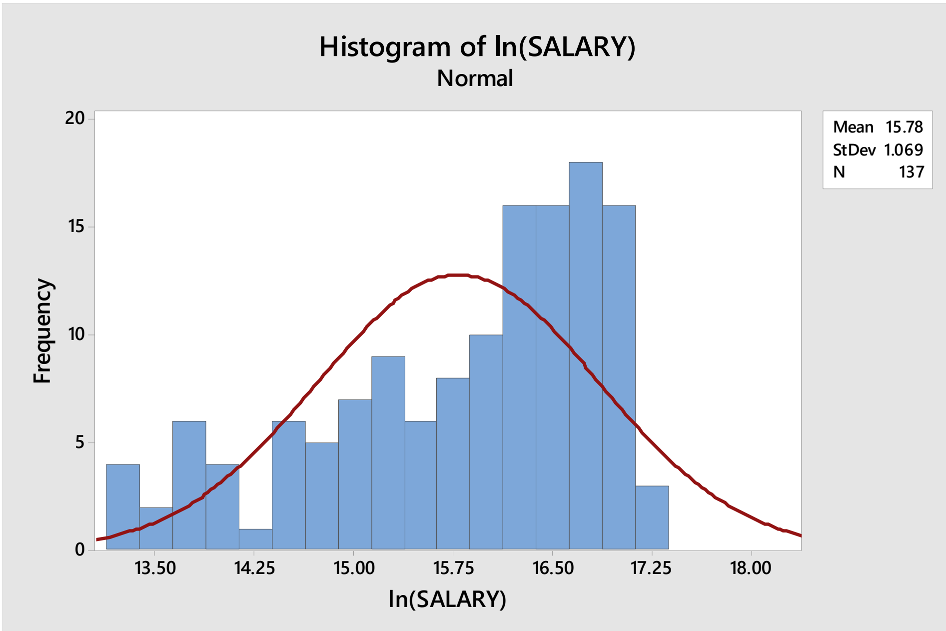

Figure 5. Histogram of ln(Salary) for Hitters Eligible for Free Agency

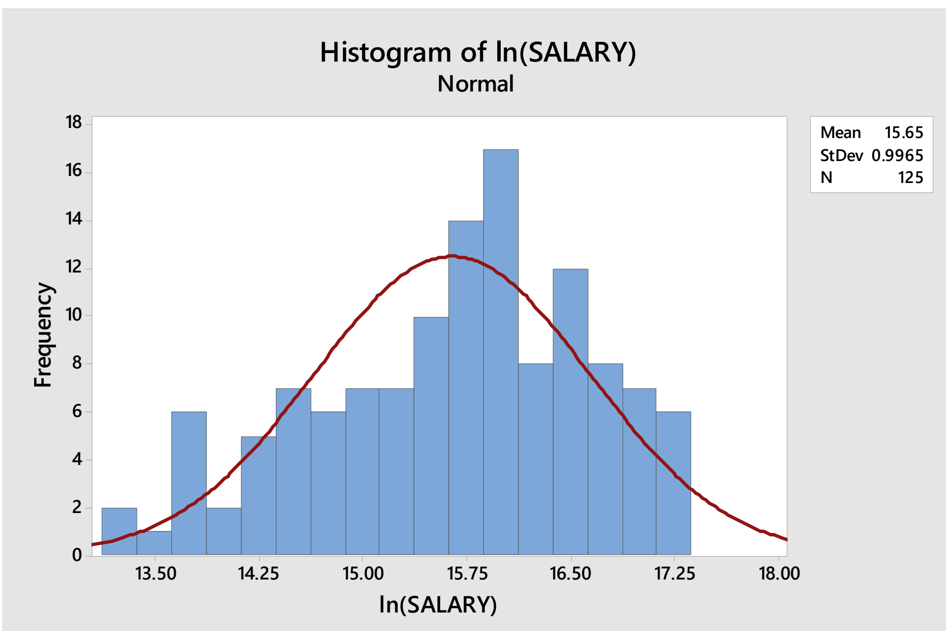

Figure 6. Histogram of ln(Salary) for Pitchers Eligible for Free Agency

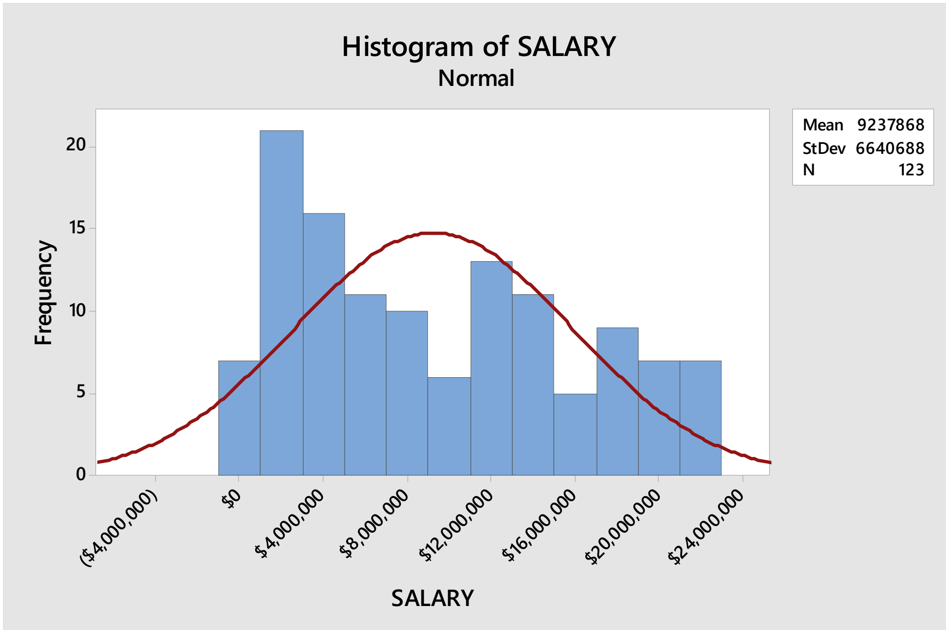

Figure 7. Histogram of Salary for Bottom 90% of Hitters Eligible for Free Agency

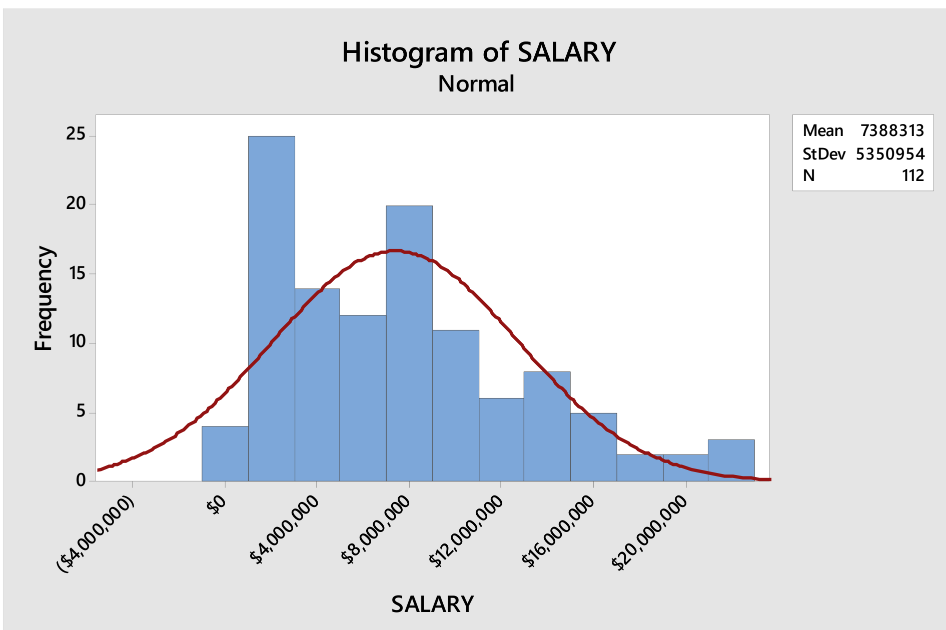

Figure 8. Histogram of Salary for Bottom 90% of Pitchers Eligible for Free Agency

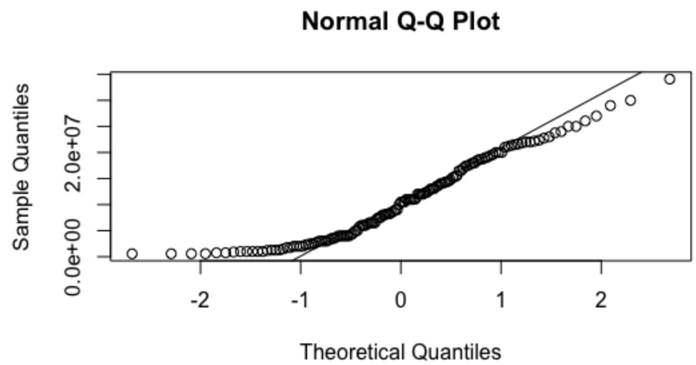

Figure 9. QQ-Plot of Salary for Hitters Eligible for Free Agency

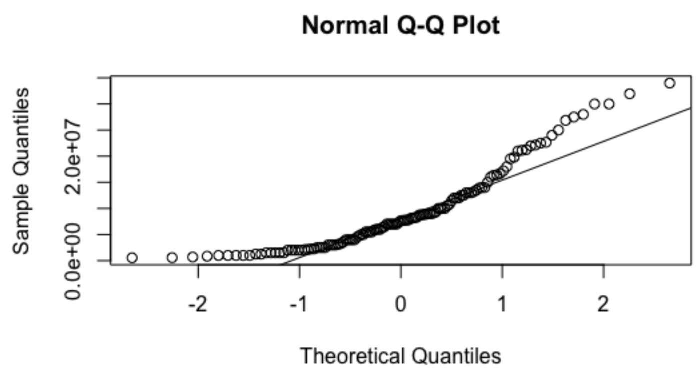

Figure 10. QQ-Plot of Salary for Pitchers Eligible for Free Agency

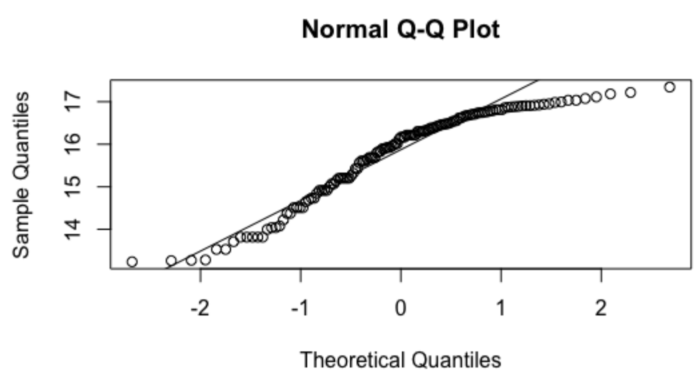

Figure 11. QQ-Plot of ln(Salary) for Hitters Eligible for Free Agency

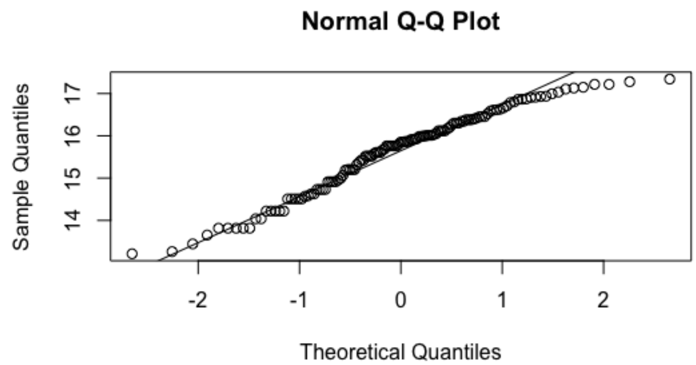

Figure 12. QQ-Plot of ln(Salary) for Pitchers Eligible for Free Agency

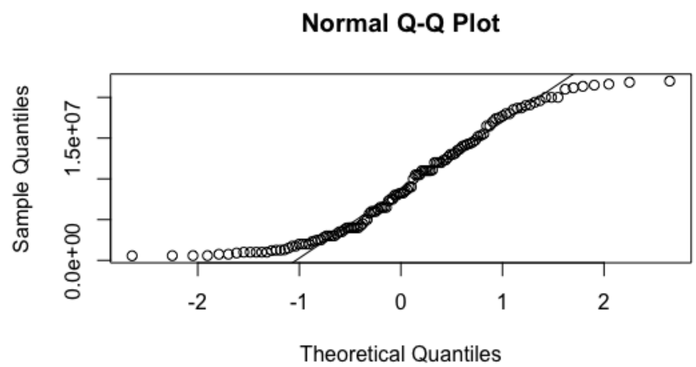

Figure 13. QQ-Plot of Salary for Bottom 90% of Hitters Eligible for Free Agency

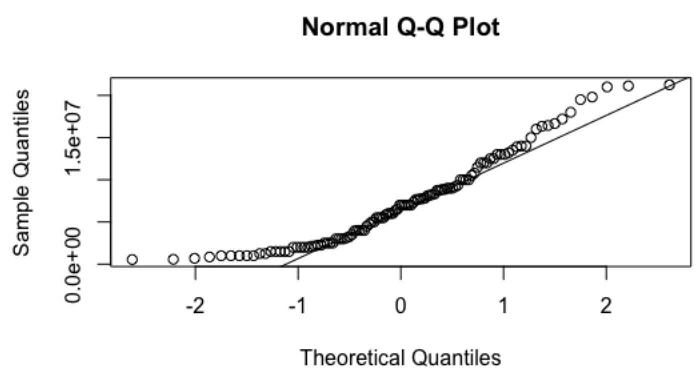

Figure 14. QQ-Plot of Salary for Bottom 90% of Pitchers Eligible for Free Agency

Appendix C

Relevant Datasets and Code

FINAL
PITCHING.xlsx

- Dataset of all pitchers, salaries, and performance statistics

FINAL
HITTING.xlsx

- Dataset of all hitters, salaries, and performance statistics

FINAL PITCHING
FA-top10.xlsx

- Dataset of pitchers, salaries, and performance statistics for bottom 90% of pitchers

FINAL HITTING FA-
top10.xlsx

- Dataset of hitters, salaries, and performance statistics for bottom 90% of hitters

R Code
Document.docx

- All R Code used for analysis

BIBLIOGRAPHY

Berry, A. (2018, April 9). Felipe Rivero changes name to Felipe Vazquez. Retrieved November 11, 2018, from https://www.mlb.com/news/felipe-rivero-changes-name-to-felipe-vazquez-c271517836

Bradley, T. (2018, December 18). Nationals Sign Matt Adams. Retrieved from https://www.mlbtraderumors.com/matt-adams

Brown, D. T., Link, C. R., & Rubin, S. L. (2017). Moneyball After 10 Years: How Have Major League Baseball Salaries Adjusted? *Journal of Sports Economics*,*18*(8), 771-786. Retrieved September 12, 2018.

Collective Bargaining Agreement 2017-2021. (2016, November 30). Retrieved September 10, 2018, from http://www.mlbplayers.com/pdf9/5450407.pdf

Budd, J. W., Sojourner, A., & Jung, J. (2017). ARE VOLUNTARY AGREEMENTS BETTER? EVIDENCE FROM BASEBALL ARBITRATION. *ILR Review,70*(4), 865-893. Retrieved September 28, 2018.

Deli, D. (2012). Assessing the Relative Importance of Inputs to a Production Function. *Journal of Sports Economics,14*(2), 203-217. doi:10.1177/1527002511434056

Drenon, T. (2014, February 17). Julio Teheran extension breakdown. Retrieved from https://www.sbnation.com/mlb/2014/2/17/5418896/julio-teheran-extension-braves-contract-breakdown

Hill, J. R., & Jolly, N. A. (2017). Revenue Sharing and Player Salaries in Major League Baseball. *Journal of Sports Economics,18*(8), 831-849. Retrieved October 9, 2018.

Home - Baseball Prospectus. (n.d.). Retrieved January 25, 2019, from https://www.baseballprospectus.com/

MLB Arbitration Tracker. (n.d.). Retrieved December 28, 2018, from https://www.spotrac.com/mlb/arbitration/2018/

MLB Salaries - MLB Baseball. (n.d.). Retrieved October 10, 2018, from https://www.usatoday.com/sports/mlb/salaries/2018/player/all/

MLB Statistics - MLB Baseball. (n.d.). Retrieved October 10, 2018, from https://www.usatoday.com/sports/mlb/statistics/2017/season-regular/all/player/batting/

MLB Statistics - MLB Baseball. (n.d.). Retrieved October 10, 2018, from https://www.usatoday.com/sports/mlb/statistics/2017/season-regular/all/player/pitching/

Todd, J. (2018, November 27). Reds Agree To Three-Year Deal With Raisel Iglesias. Retrieved from https://www.mlbtraderumors.com/2018/11/reds-agree-to-three-year-deal-with-raisel-iglesias.html

Todd, J. (2018, May 21). Jorge Soler Hires ACES. Retrieved from https://www.mlbtraderumors.com/2018/05/jorge-soler-hires-aces.html

Walters, S. J., Von Allmen, P., & Krautmann, A. (2017). Risk Aversion and Wages: Evidence from the Baseball Labor Market. *Atlantic Economic Journal,45*(3), 385-397. Retrieved September 15, 2018.

Woodbery, E. (2017, December 08). Who is Mike Fiers? 6 things to know about new Tigers starter. Retrieved from https://www.mlive.com/tigers/2017/12/who_is_mike_fiers_6_things_to.html

2017 Major League Baseball Standard Batting. (n.d.). Retrieved October 10, 2018, from https://www.baseball-reference.com/leagues/MLB/2017-standard-batting.shtml

2017 Major League Baseball Standard Pitching. (n.d.). Retrieved October 10, 2018, from https://www.baseball-reference.com/leagues/MLB/2017-standard-pitching.shtml

ACADEMIC VITA

Jack Rutherford
jackrutherford53@gmail.com

Education

The Pennsylvania State University, Schreyer Honors College — **University Park, PA**
B.S. in Statistics (Applied Option), Eberly College of Science — Graduation: May 2019
B.S. in Economics (Honors Program), College of the Liberal Arts
Minor in Mathematics
Working Honors Thesis Title: *Inefficiencies in the MLB Labor Market Caused by Arbitration*

Professional Experience

Bank of America Corporation — **New York, NY**
Corporate Audit Summer Analyst — *June 2018-August 2018*
- Constructed a logistic regression model used to predict consumer vehicle lending decisions using RStudio
- Tested and documented the control points within the system used for the production of letters to consumers
- Examined crucial systems used in the aggregation and reporting of market risk data

The Pennsylvania State University — **University Park, PA**
Research Experience for Undergraduates Program, Economics Dept. — *January 2018-May 2018*
- Collaborated with a professor in research and econometric modeling
- Utilized principles of time series and multiple linear regression using economic data to build models in EViews
- Substantiated macroeconomic theories using model results

The Pennsylvania State University — **University Park, PA**
Undergraduate Teaching/Grading Assistant — *August 2016-May 2018*
- Assisted professors in teaching an introductory microeconomics course and an introductory sociology course
- Managed the grading of papers as well as in-class assignments
- Conducted review sessions prior to exams/quizzes and proctored exams

Italian Center of Stamford, Inc. — **Stamford, CT**
Accounting Office Intern, Food Services — *May 2017-August 2017*
- Documented significant loss of income due to a fire for an insurance claim using Microsoft Excel
- Corresponded with customers to schedule events on the grounds
- Managed service to guests in the snack bar and at scheduled events

Leadership and Team Experience

Phi Gamma Delta Fraternity — **University Park, PA**
Historian, Risk Management Chair, Judicial Board Member — *September 2016- Present*
- Elected by the brotherhood to serve as a member of the house executive board
- Manage compliance to University risk protocols at all chapter events

Computer Skills

- Word
- Excel
- PowerPoint
- Minitab
- R, RStudio
- SAS
- Stata
- EViews
- Python