Predicting Average Annual Value of Free Agent Contracts in Major League Baseball

by

Anton Libsch

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Applied Statistics

by

May 2018

APPROVED:

Professor Thelge Buddika Peiris, Major Thesis Advisor

Abstract

This project uses multiple linear regression to predict the value of Major League Baseball free agent contracts, inspired by the low volume of published research on this topic. I found one published paper that shared my research goal but its predictive power was in need of improvement. An in depth comparison of our models is carried out with k-fold cross validation mean square prediction error being used as the main standard. The predictor variables considered in my models were related to performance evaluation and position, and the response variable was inflation-adjusted average annual value of the contract. The result of the project is two linear regression models, one for hitters and one for pitchers.

Acknowledgements

Special thanks to those that contributed in making my completion of this thesis possible.

Prof. Peiris, who served multiple roles through this process. As Professor of my linear regression course, and as both my academic and thesis advisor. His guidance and patience was invaluable throughout the process of completing this project.

My parents and family, for teaching me to always strive to reach my full potential as a student and an individual, equipping me to complete my degree at WPI.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Since professional baseball was established in the United States, valuation methods of player performance used by baseball teams and fans have continued to evolve. In 1859, Henry Chadwick introduced the box score, which enabled the development of simple, individual performance metrics. A century later, statisticians were applying advanced statistical tools to formulate advanced, empirical performance metrics, an application known as sabermetrics. Today, sabermetrics often includes technology to more accurately measure player contribution. Although the field of baseball analytics has seen incredible progress, the work referenced above has been made by independent parties. Their work had not penetrated Major League Baseball (MLB) front offices until only a couple of decades ago, leaving MLB teams to use outdated, simplistic metrics in evaluating the value of players' contributions.

The book *Moneyball: The Art of Winning an Unfair Game*, written by Michael Lewis, describes the Oakland Athletics' path to being the first team known to apply some of these theories to player valuation, and their resulting qualification for the playoffs in 2002 despite having the 3rd lowest payroll in the MLB. Their forward thinking front office's acknowledgement of advanced analytical theories put them ahead of their time, enabling them to find and acquire effective players who were undervalued by the rest of the league. Since then, many teams have implemented sabermetrics, but to varying degrees. As a result, different teams have different criteria in their valuation of players.

Professional baseball teams seek to make profits. Teams sign players to contracts, paying them to play, in turn generating revenue to continue the cycle. When teams are made up of better players, they tend to win more games, generating even more revenue. In this sense, paying for good players is a good investment for a team to make. This leads to the question, “How does a team determine what a specific player is worth?” How much a team values a player can be observed by how much it is willing to pay him.

Teams' differing approaches to evaluating player performance can lead them to different conclusions about how much they should pay any given player. Given this source of variation around a player's worth, there is constant discussion amongst observers and insiders of the sport as to how much a player can expect to be paid, or how much money a player is worth.

There are not many publically available studies that use regression analysis to attempt to answer this question. I found two that were noteworthy. The first was published in 2013 by Tyler Wasserman and uses a variety of factors, including one performance metric, to predict player salary. The other study was published in 2015 by Rhonda Magel and Michael Hoffman, and uses various performance metrics to predict salary.

The purpose of my thesis is to answer the question of which performance metrics are most predictive of player value to MLB teams. Despite the aforementioned works, this thesis fills a gap in knowledge that currently exists among people who have an interest in baseball and statistics. Most factors used in Wasserman's study are not performance metrics, so it provides a broader explanation of player salary. Magel and Hoffman's study has the same goal as I have for my thesis, so in this paper I will attempt to present a model with superior predictive power to theirs.

Using linear regression, I predicted the average annual value (AAV) of MLB free agent contracts signed leading up to the 2015-2017 seasons. Since the purpose of my study concerns player performance, the predictor variables for my models were strictly performance metrics and binary variables explaining player positions and roles.

This remainder of this thesis is organized as follows. Chapter 2 includes contextual information that is helpful in understanding this paper and its significance. Chapter 3 includes the theory behind my methods of fulfilling the goal of this project as well as the preparation for my analysis. In chapter 4, I will explain the results of my analysis. The significance of my results and potential future related works are discussed in chapter 5.

Chapter 2

Background

In this chapter I will discuss the context of the question that my research answers. This includes concepts that the reader should be familiar with in order to best understand the analysis, and similar research that has been published.

2.1 Free Agency

The significance of a player becoming a free agent is that he becomes eligible to negotiate a new contract with any team. This is an important concept in the context of my research because when teams have to bid against each other, the player is most likely to receive at least one offer that is closest to his expected worth. Major League Baseball's Collective Bargaining Agreement (CBA) outlines the standard procedure for a player to qualify as a free agent, which is to meet the following conditions: (1) expiry of previous contract and (2) accumulation of six years of service time.

Article XX of the 2017-2021 MLB Collective Bargaining Agreement (CBA) states “Following the completion of the term of his Uniform Player's Contract, any Player with 6 or more years of Major League service who has not executed a contract for the next succeeding season shall become a free agent, subject to and in accordance with the provisions of this Section B."

Article XXI of the 2017-2021 MLB Collective Bargaining Agreement states "One full day of Major League service will be credited for each day of the championship season a Player is on a Major League Club's Active List. A total of 172 days of Major League credited service will constitute one full year of credited service. A Player may not be credited with more than one year of credited service, 172 days, in one championship season (including any tie-breaker games)." It is worth noting that a player continues to accumulate service while on the disabled list.

A player who has not met the above conditions can be granted free agency by his team through two methods. (1) A team can a release a player who is under contract, which is essentially a contract termination. (2) Note that according to Article XX of the CBA, a player without a contract for the following season only becomes a free agent if he has accumulated six years of service. This is because if a player is without a contract for the next season but has accumulated less than six years of service, he is only allowed to negotiate a new contract with his current team. In the event that a team does not wish to re-sign such a player, it can relinquish its exclusive negotiation rights with the player, making him a free agent. This is known as a decision to non-tender a player.

2.2 Previous Research

There is limited complete, published research that implements regression analysis to predict MLB player salaries. Two that I found noteworthy are described below.

The first published research of note is titled *Determinants of Major League Baseball Player Salaries* (Wasserman, 2013). Its goal was similar to mine in the sense that it attempts to predict player contract value, but there are a couple of significant differences. One is that the model's observations are not limited to free agent signings. I explain in section (3.3) why this is a potentially major difference. The most significant difference in our models is our predictors of choice. Where I am only concerned with performance metrics, he is concerned with “a multitude of different factors which may have significant relationships to salaries” (Wasserman, 2013). Due to his broader interest in predictors of salary, his model includes player performance, age, agent, team, timing of the signing, inclusion of no-trade clause, and years remaining on the contract prior to signing. These differences show that this research does not attempt to fill the gap in knowledge that I sought to fill.

The second published research of note is titled *Predicting Salaries of Major League Baseball Players* (Magel and Hoffman, 2015). This had the same goal as I did, to find a statistical relation between player performance metrics and player salaries. They developed two models each for hitters and pitchers: one based on yearly data and the other on career data. For each of these models, the response variable was the natural log of player salary, and the predictor variables consisted of performance metrics and positional/contextual data.

Despite the fact that our research had the same goal, there are some aspects to the research that could be improved upon to yield a superior model. The final yearly models contained some insignificant predictors, were a weak fit to the data, and whether it met the assumptions of the linear regression model was not stated.

Chapter 3

Methodology

In this chapter I will explain the steps that I took leading up to attaining my results. This includes the pertinent basic theoretical elements of the method that I used to generate my models, linear regression. This chapter also includes details on the composition of my dataset, including models and variables used.

3.1 Regression Analysis

Regression analysis is a statistical methodology that utilizes the relation between two or more quantitative variables so that a response or outcome variable can be predicted from the other, or others (Kutner 2004). Regression analysis is often applied to problems where the change in the response variable in relation to the predictor variables is of interest.

3.1.1 Relations Between Variables

The implementation of regression analysis to create a model where one linear predictor variable is used to predict a linear response variable is called simple linear regression.

Simple linear regression is used when the change in the response variable can be explained by a single predictor variable.

In some cases where regression analysis can be applied, there are multiple variables that significantly affect the response variable. In such cases, a model containing a single predictor variable may provide an inadequate description of the response variable, hence be too imprecise to be useful. To provide more precise predictions of the response variable, one would need a model containing additional predictor variables. When regression analysis is applied in this way - with multiple predictor variables, all linear, the application is known as multiple linear regression.

3.1.2 Purpose of Regression Analysis

Examples of regression analysis being applied are: (1) A high school principle wants to predict exam scores, so the relationship between studying and exam score is utilized. (2) Actuaries at a life insurance company want to predict life expectancy, so the relationship between smoking frequency, blood pressure, and age at death is utilized.

In the first example, the principle uses simple linear regression. Hours spent studying is the predictor variable, and exam score is the response variable. In the second example, the actuaries use multiple linear regression. Number of cigarettes smoked per day is the first predictor variable, systolic blood pressure is the second, and diastolic blood pressure is the third. Age at death is the response variable.

Regression analysis serves three main purposes: (1) Description: Understanding how certain factors influence a response variable, (2) Control: Altering the value of a response variable through the manipulation of certain factors, (3) Prediction: Forecasting a response variable based on its statistical relation with certain predictor variables.

It is common for the purposes of regression analysis to overlap in practice. Consider the example of the high school principle. With the goal of increasing student performance on exams, the principle could use regression analysis so that teachers will better understand the impact of studying, and encourage their students to do so accordingly. If it is known that incoming students face extenuating circumstances that make studying difficult, the principle could make adaptations in the school to accommodate the their needs. Here, all three main purposes are of value.

3.1.3 A Note on Causality

The presence of a statistical relation between a response variable and one or more explanatory variables does not imply that a causal relationship exists.

Consider an example where regression analysis is applied to use vocabulary size to predict writing speed among young children. A positive relation would not imply that an increase in vocabulary size *causes* an increase in writing speed. We know that there are other variables not included in the model that are positively correlated with both vocabulary size and writing speed, such as education level and age.

Consider another example where regression analysis is applied to use the reading on a scale to predict a persons body mass. Although there would be a positive relation, we know that an increase in the reading on a scale does not cause an increase in a person's body mass. In reality, increased body mass would cause the reading on a scale to be higher.

Therefore, it is important to be cautious in drawing conclusions about causality from regression analysis. Further analyses would be necessary to draw conclusions about the causality between variables.

3.1.4 The Regression Models

Simple Linear Regression Model

Consider the case where there is one predictor variable, X. The simple linear regression model is stated as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, ..., n, \tag{3.1}$$

where Y_i is the value of the response variable for the i^{th} observation; β_0 and β_1 are parameters; X_i is a known constant, the value of the predictor variable for the ith observation; and ε_i is a random error term.

This model is fitted under the following assumptions: (1) The error term has a mean of 0, $E\{\varepsilon_i\} = 0$, $i = 1, ..., n$ (2) The error term has a constant variance, $\sigma^2\{\varepsilon_i\} = \sigma^2$, $i = 1, ..., n$ (3) The error terms are uncorrelated, $\sigma\{\varepsilon_i, \varepsilon_j\} = 0$ for all $i, j = 1, ..., n; i \neq j$.

The parameters β_0 and β_1 are called regression coefficients. β_1 represents the change in the mean value of Y per unit increase in X. β_0 represents the mean value of Y when $X = 0$, assuming that $X = 0$ is in the scope of the model.

The response variable of the regression model, Y_i, is the sum of two components: (1) the constant term $\beta_0 + \beta_1 X_i$ and (2) the random term ε_i. The inclusion of the second component means that Y_i is a random variable, so its expectation and variance can be calculated. Keeping in mind that its first component is a constant and its second component, ε_i, satisfies $E\{\varepsilon_i\} = 0$ and $\sigma^2\{\varepsilon_i\} = \sigma^2$, we see that:

$$\begin{aligned} E\{Y_i\} &= E\{\beta_0 + \beta_1 X_i + \varepsilon_i\}, \\ &= E\{\beta_0 + \beta_1 X_i\} + E\{\varepsilon_i\}, \end{aligned}$$

$$= \beta_0 + \beta_1 X_i + 0.$$

and

$$\begin{aligned}\sigma^2\{Y_i\} &= \sigma^2\{\beta_0 + \beta_1 X_i + \varepsilon_i\}, \\ &= \sigma^2\{\beta_0 + \beta_1 X_i\} + \sigma^2\{\varepsilon_i\}, \\ &= 0 + \sigma^2.\end{aligned}$$

Here, we are shown that the response Y_i comes from a probability distribution whose mean and variance are:

$$E\{Y_i\} = \beta_0 + \beta_1 X_i, \text{ and } \sigma^2\{Y_i\} = \sigma^2 \tag{3.2}$$

It follows that the response function for regression model (3.1) is

$$E\{Y\} = \beta_0 + \beta_1 X. \tag{3.3}$$

Multiple Linear Regression Model

I Consider the case where there are $p - 1$ predictor variables, $X_1, ..., X_{p-1}$. The regression model is stated as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \qquad i = 1, ..., n \tag{3.4}$$

It can also be expressed as

$$Y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \varepsilon_i \tag{3.5}$$

Similar to the simple linear regression model, Y_i is the value of the response variable for the ith observation; β_k, $k = 0, ..., p-1$ are parameters; X_{ik}, $k = 0, ..., p-1$ are known constants, the values of the predictor variables for the i^{th} observation; ε_i is a random error term.

This model is fitted under the following assmptions: (1) The error term has a mean of 0, $E\{\varepsilon_i\} = 0$, $i = 1, ..., n$ (2) The error term has a constant variance, $\sigma^2\{\varepsilon_i\} = \sigma^2$, $i = 1, ..., n$ (3) The error terms are independent, $\sigma\{\varepsilon_i, \varepsilon_j\} = 0$ for all $i, j = 1, ..., n; i \neq j$.

The parameters β_k, $k = 0, ..., p-1$ are the regression coefficients. For each value of $k = 0, ..., p-1$, β_k represents the change in the mean value of Y per unit increase in X_k when the other predictor variables are held constant. β_0 represents the mean value of Y when $X = 0$, assuming that $X = 0$ is in the scope of the model.

Following the same reasoning as in simple linear regression, we take the expectation and variance of Y_i from equation (3.5) and see that:

$$\begin{aligned}
E\{Y_i\} &= E\{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \varepsilon_i\}, \\
&= E\{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1}\} + E\{\varepsilon_i\}, \\
&= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1}.
\end{aligned}$$

and

$$\begin{aligned}
\sigma^2\{Y_i\} &= \sigma^2\{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \varepsilon_i\}, \\
&= \sigma^2\{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1}\} + \sigma^2\{\varepsilon_i\}, \\
&= \sigma^2.
\end{aligned}$$

Here, we are shown that the response Y_i comes from a probability distribution whose mean iand variance are:

$$E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1}, \text{ and } \sigma^2\{Y_i\} = \sigma^2 \tag{3.6}$$

It follows that the response function for regression model (3.5) is

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{p-1} X_{p-1}. \tag{3.7}$$

3.1.5 Linear Regression with Qualitative Variables

Regression analysis can be applied to problems that include not only quantitative variables, but also qualitative variables, such as sex (male, female) or disability status (not disabled, partially disabled, fully disabled). We refer to the values that a qualitative variable could take on as classes. When fitting a regression model to a problem that includes a qualitative variable with c classes, we use $c - 1$ indicator variables that take on the values 0 and 1 to represent them.

An indicator variable takes on the value 1 for an observation that is of its corresponding class. An indicator variable takes on the value 0 for an observation that is not of its corresponding class. If an observation is of the c^{th} class, the class without a corresponding indicator variable, we let each indicator variable in the model take on the value 0.

The coefficients of these indicator variables measure the differing effects of the qualitative variable classes on the mean of the response function compared to the class for which there is no indicator variable.

Consider an example where regression analysis is used to predict the length of hospital stay (Y) based on the patient's age (X_1) and sex (X_2). Where sex has classes male and female, the regression model for this example is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \qquad i = 1, ..., n \tag{3.8}$$

where (X_1) takes on the value of the patient's age, (X_2) takes on the value 1 if the patient is a male and 0 if the patient is a female.

Consider the same example, but the hospital decides to add disability status as another predictor variable. Where disability status has classes not disabled, partially disabled, and fully disabled, the regression model for this example is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i, \qquad i = 1, ..., n \tag{3.9}$$

where X_1 takes on the value of the patient's age, X_2 takes on the value 1 if the patient is a male and 0 if the patient is a female, X_3 takes on the value 1 if the patient is not disabled and 0 otherwise, and X_4 takes on the value 1 if the patient is partially disabled and 0 otherwise. A fully disabled patient would be represented by both X_3 and X_4 taking on the value 0.

3.2 Models & Variables

In this section I explain decisions that I made to increase the predictive power and meaningfulness of my regression models and their variables. Some adjustments were made to my response variable in order to increase its practical interpretation. After fitting a model for all hitters and a model for all pitchers, it became apparent that predictive power could be increased for pitchers if they are separated into two models: one for starting pitchers and another for relief and closing pitchers.

3.2.1 Response Variable

Interest Rates

The value of money depends on the time of its receipt, due to the fact that money is capable of earning interest. A fixed amount of money is worth more the sooner it is received, because the recipient gains the opportunity to invest it, increasing its value. Given an interest rate, one can calculate the present value of a fixed dollar amount from a different year, enabling accurate comparison of dollar totals from different years. Since I am comparing payments made in different years, it is imperative that I account for this changing value of money over time. In the MLB, the changing value of a fixed dollar amount that a team pays towards a player's contract is determined by the change in funds that it has dedicated to player contracts. For this problem, this change in funds will estimate interest rate in present value calculations. I expect that this change would be influenced by many factors, some of which are unpredictable and not equally influential in each year, so having a different interest rate for each year is expected. As an observer of the sport, I am unable to determine the exact amount of money MLB teams dedicate to player salaries each year. The most effective way to approximate this total was to observe the total amount that MLB teams spent on player salaries each year. I obtained a unique interest rate value for each year by dividing its total MLB payroll by that of the previous season. The results are shown in the table below, where (i_{year}) represents the change in total MLB salaries from the previous year to the year indicated in the subscript.

Table 3.1: Interest Values

Year	Total MLB Salaries	Change from Previous Year (i_{year})
2017	$\$4,566,641,086$	1.0495
2016	$\$4,351,286,371$	1.0636
2015	$\$4,091,061,330$	1.1009
2014	$\$3,716,027,120$	1.0836
2013	$\$3,429,276,085$	

The 2017 value of a salary, $\$x$, paid in 2015 can be shown as

$$\begin{aligned} 2017Value &= \$x \times i_{2016} \times i_{2017}, \\ &= \$x \times 1.0636 \times 1.0495. \end{aligned}$$

However, this method would not work for salaries after 2017, since the total MLB payroll for these years was not yet known. To get the interest rates for years after 2017 I took a weighted average of the change in total MLB salaries from 2013-2017 to project the future change in total MLB salaries, as follows:

$$i_{projected} = \frac{4 \times 1.0495 + 3 \times 1.0636 + 2 \times 1.1009 + 1 \times 1.0836}{10} = 1.0674 \qquad (3.10)$$

When applying this to my dataset, I assumed that all contracts were paid out at the end of the season, unless the players contract stated otherwise.

Composition of the Response Variable

Average Annual Value weighted by inflation (AAV) is the response variable in my models. For the remainder of this paper, when I refer to AAV, it should be assumed that it is weighted by inflation. I chose AAV rather than total contract value because (1) the latter is heavily affected by contract length, which in turn is heavily affected by the player's age and (2) AAV shows a player's perceived value in a single season.

Prior to constructing AAV, each observation in the datasets had an entry for each salary payment of its contract. The first step to constructing AAV was to convert all contract payments (salaries and signing bonuses) to their 2017 values for each observation. I made the assumption that all contract payments are made at the end of the season indicated in the contract. The second step was to calculate the sum of all contract payments for each observation, resulting in the 2017 value of the total contract. The third step was to divide this sum by the number of guaranteed years in the contract for each observation. These values are all in millions of dollars.

MLB contracts are known to be very complex. This made it difficult to decide the best way to value contract. For simplicity and consistency, I decided to only include guaranteed money. Some relatively common contract clauses and other factors that were ignored in the AAV calculation are: opt-out clauses, performance bonuses, trade incentives, non-monetary perks (special hotel accommodations, etc.), no-trade clauses, and rejection of a qualifying offer.

3.2.2 Predictor Variables

The variables that were considered for each model prior to selecting the best subset will be shown in this section. Hitter and pitcher AAV were predicted in separate models due to their differing performance metrics, hence their differing predictor variables.

A variable ending with .1 represents a performance metric from the season prior to the signing of the contract. A variable ending with .2 represents a performance metric from two season prior to the signing of the contract.

Model for Hitters

Below are the predictor variables that were considered for the Hitters Model with the name of the performance metric that is being used.

Age — The players age at the signing of the contract

PA.1 & PA.2 — Plate Appearances

R.1 & R.2 — Runs scored

HR.1 & HR.2 — Home Runs

RBI.1 & RBI.2 — Runs Batted In

SB.1 & SB.2 — Stolen Bases

AVG.1 & AVG.2 — Batting Average

OBP.1 & OBP.2 — On Base Percentage

SLG.1 & SLG.2 — Slugging Percentage

wOBA.1 & wOBA.2 — Weighted On Base Average

wRCp.1 & wRCp.2 — Weighted Runs Created Plus

wRAA.1 & wRAA.2 — Weighted Runs Above Average WAR.1 & WAR.2 — Fangraphs Wins Above Replacement

Bat.1 & Bat.2 — Batting Runs (A Fangraphs Metric)

BsR.1 & BsR.2 — Base Running (A Fangraphs Metric)

Off.1 & Off.2 — Offensive Runs Above Average (A Fangraphs Metric)

Def.Tot — Defensive Runs Above Average (a sum of Def.1 and Def.2)

PosDH — Indicator variable for the Designated Hitter position

PosC — Indicator variable for the Catcher position

Pos1B — Indicator variable for the 1st base position

Pos2B — Indicator variable for the 2nd base position

Pos3B — Indicator variable for the 3rd base position

PosSS — Indicator variable for the short stop position

An indicator variable for the outfield position was excluded.

Model for Pitchers

Below are the predictor variables that were considered for the Pitchers Model with the name of the performance metric that is being used.

Age — The players age at the signing of the contract

IP.1 & IP.2 — Innings Pitched

GS.1 & GS.2 — Games Started

CG.1 & CG.2 — Complete Games

W.1 & W.2 — Wins

BB.1 & BB.2 — Bases on Balls

SV.1 & SV.2 — Saves

HLD.1 & HLD.2 — Holds

WHIP.1 & WHIP.2 — Walks plus Hits per Innings Pitched

ERA.1 & ERA.2 — Earned Run Average

FIP.1 & FIP.2 — Fielding Independent Pitching

xFIP.1 & xFIP.2 — Expected Fielding Independent Pitching

K9.1 & K9.2 — Strikeouts per 9 Innings

BB9.1 & BB9.2 — Bases on Balls per 9 innings

HR9.1 & HR9.2 — Home Runs per 9 innings

KBB.1 & KBB.2 — Strikeouts per Base on Balls

WAR.1 & WAR.2 — Fangraphs Wins Above Replacement

xFIPm.1 & xFIPm.2 — Expected Fielding Independent Pitching Minus

PosSP — Indicator variable for Starting Pitchers

PosRP — Indicator variable for Relief Pitchers

An indicator variable for closing pitchers was excluded

RHP — Indicator variable for right-handed pitchers

An indicator variable for left-handed pitchers was excluded

3.3 Criteria for Observation Inclusion

In order to ensure that the information in my response variable was captured by my predictor variables, I placed some restrictions on which observations would be included in my models.

As mentioned in section (2.1), the significance of free agency is a player's ability to sign a new contract with any team, so only players who signed new contracts as free agents were included in my models. A player can sign a new contract outside of free agency by either signing the contract (1) before the completion of his active contract, which is known as a contract extension or (2) after the completion of his active contract if the player has less than six years of service, which is known as being tendered a contract. Observations with contract extensions were not considered for my model, as such contracts are often times worth less money than what the player could have received if every team was able to place a bid. Players often sign contract extensions because it provides security and they value playing for their current team more than maximizing their salary. Observations with tendered contracts were not considered for my model, as such players are only able to negotiate with their current team. Tendered contracts pay less than free agency contracts for players of equal ability, mainly due to a process called arbitration, which I will not go into detail on in this paper. Including observations with contract extensions or tendered contracts could have potentially introduced variation in my response variable that is not captured by my predictor variables.

Players can sign major league contracts or minor league contracts. The main differences between the two are that major league contracts pay higher salaries, and teams are highly incentivized to keep players signed to such contracts on the major league roster. To be considered for this model, a player had to have signed a major league contract as a free agent. Players who signed minor league contracts were excluded because they are often signed as extra players, unlikely to appear in many major league games, and are typically similar in talent to each other. Therefore, teams are unlikely to compete with each other to sign them.

There was a minimum playing time requirement for players to be included in my models. For the pitchers models, all starting pitchers had to have pitched in at least 20 games during the most recent season, and in at least 17 games during the previous season. All relief pitchers had to have pitched in at least 33 games during the most recent season, and in at least 24 games during the previous season. For the hitters model, all players had to have had at least 300 plate appearances during the most recent season, and at least 240 plate appearances during the previous season.

I made the decision to only include contracts signed in offseasons prior to the 2015-2017 seasons. As time progresses, teams evolve in their player evaluation techniques. Even though a larger dataset is generally better, including contracts from a larger range of seasons will likely capture more of the progression in valuation techniques, weakening the predictive power of my models.

3.4 Dataset Composition

In order to fulfill the goal of this study with regression analysis, a dataset containing the observations and their corresponding values for each predictor variable and the response variable was required. This section describes this process. Below are the steps that I took for each of the 2015, 2016, and 2017 seasons.

My first step was to obtain free agent signing data. I exported a .csv formatted dataset [baseball-reference.com] of all free agent players who signed new contracts for a given year to Microsoft Excel. At this point in the process, there were three datasets in total. The observations for my models would ultimately be the entries from these datasets that met the conditions explained later in this chapter.

My second step was to split each dataset into two: one for pitchers and one for hitters. I first had to define each player in the datasets as a hitter or a pitcher. These datasets contained basic information on the free agency signings and some simple career performance data as variables. Among these variables were "At Bats" and "Innings Pitched," which are hitters' and pitchers' playing time metrics, respectively. Whether a player was a hitter or pitcher was not specified, so I defined pitchers as any player who had more innings pitched than at bats over the course of his career, and I defined all other players as hitters.

My third step was to obtain player performance metrics data. I exported a .csv formatted dataset [fangraphs.com] of all MLB pitchers' performance metrics and a dataset of all MLB hitters' performance metrics to Microsoft Excel. The performance metrics included were those that I would consider for my model as well as those that were used in the comparison model, outlined in section (3.2).

My fourth step was to match the observations with the predictor variables. I applied the performance metrics data to their corresponding observations in the free agent signings dataset, such that a free agent signings dataset for a given year would contain performance metrics data corresponding to the each of the previous two years. I chose to do this because a team can only consider historical data when evaluating a player's performance. Positional/role related data was taken from Cot's

Contracts [baseballprospectus.com] opening day salaries spreadsheets, which contain data on players' position (hitters) and handedness (pitchers) in .csv format. Position and handedness were matched to the players in the main spreadsheet.

Most players appear at the same position for every game, but a considerable number appear at more than one position over the course of a season, let alone over the course of his career. Due to this complexity, there were cases where the Cot's Contracts spreadsheet had multiple positions listed for single players. Where this was the case, I consulted baseball-reference.com player pages where I could view how many games a player played at each position per season. I assigned these cases to the position that he played the most games in during the first year of his newly signed contract, unlike the approach that I took with performance data. I found it more appropriate to consider games during the first year of a player's contract rather than games throughout his career or in the season prior to signing the contract, because I wanted this variable to represent at which position he was signed to play. Since my model explains which criteria teams use in player valuation, this detail allows my model to consider how much teams value a player being at a certain position.

It is not uncommon for a pitcher to make some appearances as a relief pitcher and others as a starting pitcher; and among relief appearances, some in the closing pitcher role. This introduces some ambiguity when determining whether a pitcher is a starting pitcher, relief pitcher, or closing pitcher. In order to be consistent when determining a pitchers role, I used the following criteria: a pitcher who made more than half of his appearances as a starting pitcher would be considered as such; a pitcher who made at least half of his appearances as a relief pitcher and recorded at most 9 saves would be considered as such; a pitcher who made at least half of his appearances as a relief pitcher and recorded at least 10 saves would be considered a closing pitcher. At the end of the fourth step, I removed observations where the player had not had sufficient playing time. This is elaborated upon further later in the paper.

My fifth step was to record salary information. Unlike the previously mentioned data, salary data for MLB players is not publically available in .csv format. Player salary data is available on baseballprospectus.com, but not in .csv format. Therefore, I had to manually record this information in each dataset. For each observation, I recorded the duration of the contract, and the amount of salary and signing bonus paid per year. At the end of the fifth step, I removed observations where a minor league contract was signed.

My sixth step was to develop the response variable, which was developed from the salary data. This process will be explained later in the paper.

My seventh step was to combine the six datasets into two. There would be one dataset for all hitter observations and one dataset for all pitcher observations.

Chapter 4

Results

In this chapter I fit each of my models with the best subset of variables, which was found using a process known as k-fold cross validation. I chose k to be 10 for each model that I fitted. With p variables being considered for a given model, cross validation finds the best subset of i variables for each $i = 1, ..., p$. The model with the lowest Mean Squared Prediction Error (MSPR) is considered the best.

When finding the best subset of i predictors with k-fold cross validation, the dataset is first divided into k subsets of approximately equal size. The 1$^{\text{st}}$ subset is held out as a validation set and the remaining $k-1$ subsets are treated as one training set. One at a time, every combination of i predictors is fitted to the training set and its MSPR is calculated on the validation set. This process is repeated k times, with a different subset being treated as the validation set each time. The result is k MSPR values which are then averaged to obtain the k-fold cross validation error estimate (CV error) for the tested subset of predictors.

MSRP is defined as

$$MSPR = \frac{\sum_{j=1}^{n^*}(Y_j - \widehat{Y_j})^2}{n^*} \tag{4.1}$$

where n^* is the number of observations in the validation set, and $\widehat{Y_j}$ and Y_j are the fitted and actual values of the j^{th} observation, respectively.

4.1 Model for Hitters

Among the predictor variables that I considered, wOBA, wRC+, wRAA, Off, and Bat are similar in how they're calculated, so it was obvious that they would be highly correlated with each other. In regression analysis, a model loses interpretability when predictor variables are highly correlated with each other, so I needed to select only one of these. Since Off was the most correlated with AAV, I included Off.1 and Off.2 for model selection.

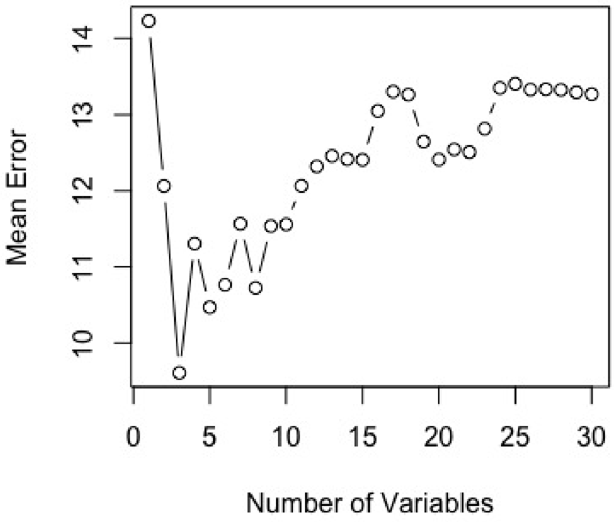

Figure 4.1: Hitters Model: CV Error

Now that I knew which variables I would consider, the next step was to perform 10-fold cross validation. The results are shown in figure (4.1). The model with the lowest CV error is that with three variables, which are WAR.1, WAR.2, and

Def.Tot. The model summary and Analysis of Variance (ANOVA) table are shown in tables (4.1) and (4.2), respectively.

Table 4.1: Hitters Model: Summary

Variable	Coefficient	t-value	Pr($>|t|$)
Intercept	1.34523	2.328	.0221
WAR.1	2.40599	12.079	<0.0001
WAR.2	1.45472	7.288	<0.0001
Def.Tot	-0.15223	-7.043	<0.0001

Table 4.2: Hitters Model: ANOVA

Variable	DF	Sum Sq	Mean Sq	F-value	Pr($> F$)
WAR.1	1	1552.82	1552.82	163.482	<0.0001
WAR.2	1	259.22	259.22	27.291	<0.0001
Def.Tot	1	471.11	471.11	49.599	<0.0001
Residuals	93	883.36	9.50		

Each predictor variable included in this model is significant, with a p-value of virtually 0, and the model has a CV error of 9.609. Other descriptive measures of this model are R^2=0.721, R^2_{adj}=0.712, AIC = 499.55, BIC = 512.42. The interpretation of the coefficients is as follows. With no change in the other predictor variables, a unit increase in WAR.1 leads to an average increase in AAV of 2.41. With no change in the other predictor variables, a unit increase in WAR.2 leads to an average increase in AAV of 1.45. With no change in the other predictor variables, a unit increase in Def.Tot leads to an average increase in AAV of -0.15, or a decrease of 0.15.

The negative coefficient for Def.Tot is surprising, given that it is a metric that measures positive contributions that a player makes to his team. To uncover the cause of the negative coefficient, I made a plot of AAV vs. Def.Tot, which can be seen in figure (4.2). This plot is supported by a correlation coefficient of -0.073. Although the correlation is weak, its negative nature is consistent with the coefficient obtained in the model.

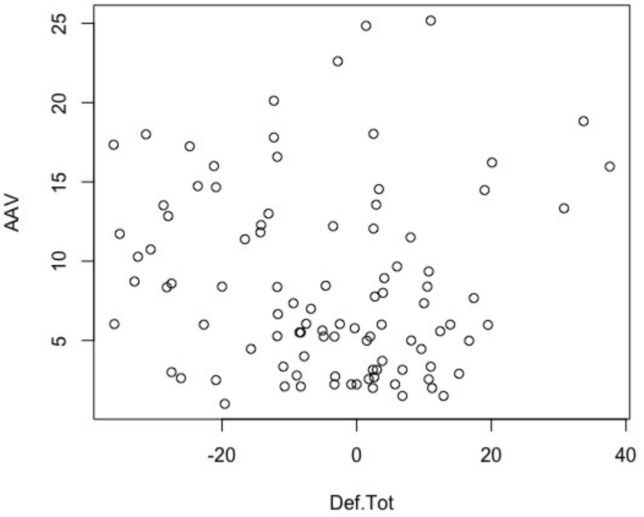

Figure 4.2: Scatterplot: AAV vs. Def.Tot

Despite the low correlation, Def.Tot is very important to the model, as removing it causes the prediction error to rise to 15.268. The ANOVA table shows that Def.Tot has an extra sum of squares value of 471.11, meaning that even after adding WAR.1 and WAR.2 to the model, Def.Tot still accounts for a lot of the variation in AAV.

Statistical studies show that simply due to the nature of baseball, there is more potential to add value to one's team on offense than on defense. It seems that although player's defensive contributions are important, teams do not mind paying more money to players with poor defensive skills as long as they can make up for their defensive deficiencies by performing well on offense.

It is important to check that my model meets the assumptions stated in section (3.1.4). I obtained a scatterplot of the model's residuals versus its fitted values, and a normal probability plot to visually check the assumptions of a multiple linear regression model. The residuals appear to be randomly distributed around the horizontal line through the origin, and appear to be approximately normally distributed.

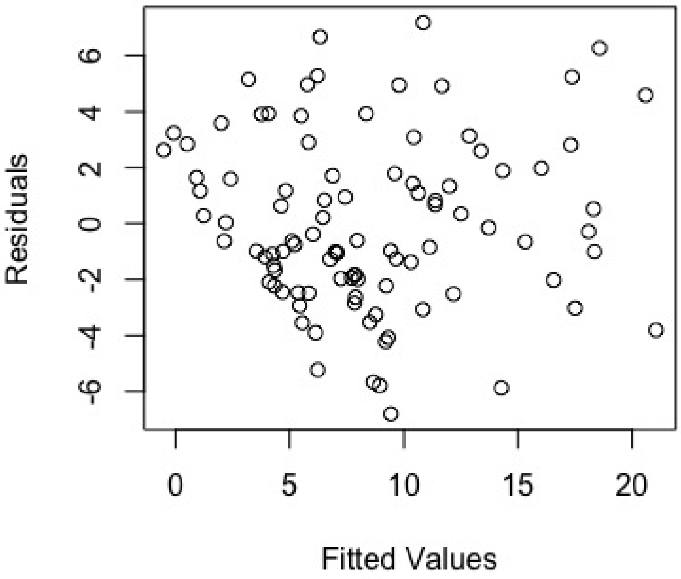

Figure 4.3: Hitters Model: Residuals Plot

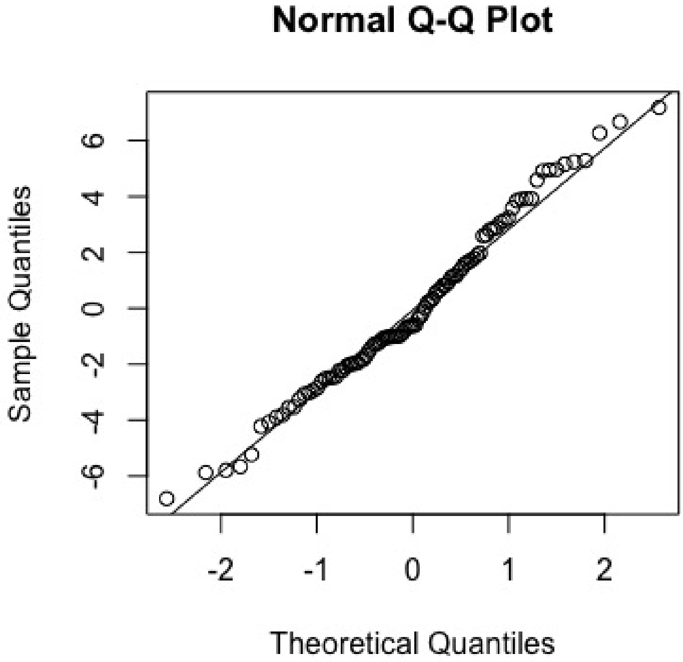

Figure 4.4: Hitters Model: Normal Probability Plot

Since the model meets the assumptions, linear regression is appropriate for the fitted data, yielding the regression function

$$\text{AAV} = 1.34523 + 2.40599 \times \text{WAR.1} + 1.45472 \times \text{WAR.2} - 0.15223 \times \text{Def.Tot.} \quad (4.2)$$

4.2 Model for Pitchers

The 10-fold cross validation results are shown in figure (4.5). The model with the lowest CV error uses seven predictor variables, which are: Age, WAR.1, WAR.2, WHIP.1, BB9.1, HR9.1, and PosRP. Since PosRP is indicator variable of a qualitative variable made up of three classes, the other indicator variable, PosSP, needs to be included with it. This yields a model with eight variables, of which the summary (table 4.3) and ANOVA table (table 4.4) are below.

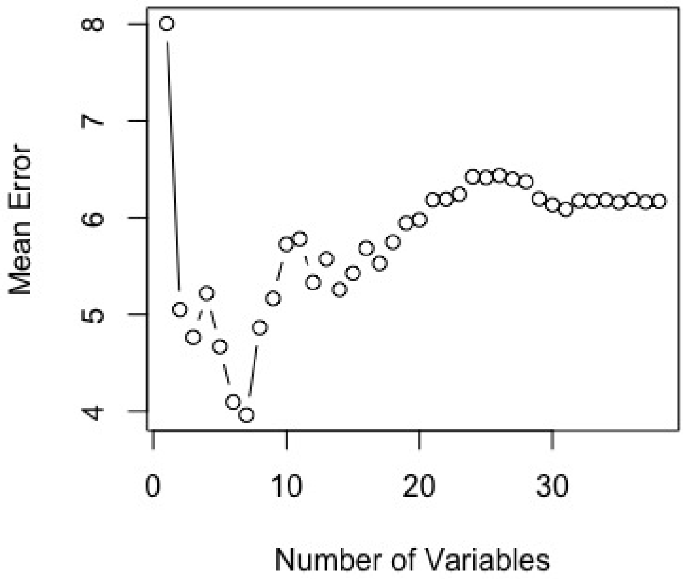

Figure 4.5: Pitchers Model: CV Error

Table 4.3: Pitchers Model: Summary (Incorrect Signs)

Variable	Coefficient	t-value	Pr($>|t|$)
Intercept	12.09753	4.987	<0.0001
Age	-0.18789	-3.635	0.0004
WAR.1	2.62371	10.077	<0.0001
WAR.2	1.37199	7.366	<0.0001
WHIP.1	-5.06653	-4.171	<0.0001
BB9.1	0.89991	3.541	0.0006
HR9.1	1.35835	2.435	0.0167
PosRP	-1.87452	-3.639	0.0004
PosSP	-0.34338	-0.521	0.6033

Table 4.4: Pitchers Model: ANOVA (Incorrect Signs)

Variable	DF	Sum Sq	Mean Sq	F-value	Pr($> F$)
WAR.1	1	2836.27	2836.27	840.5972	<0.0001
WAR.2	1	298.64	298.64	88.5087	<0.0001
Age	1	26.86	26.86	7.9602	0.005758
WHIP.1	1	25.00	25.00	7.4084	0.007648
PosRP	1	72.89	72.89	21.6030	<0.0001
PosSP	1	0.52	0.52	0.1543	0.695303
BB9.1	1	26.24	26.24	7.7765	0.006327
HR9.1	1	20.00	20.00	5.9277	0.016659
Residuals	101	340.79	3.37		

Before analyzing the fit of this model, I noticed that the coefficients for BB9.1

and HR9.1 are positive, even though a lower value for these metrics indicated a better performance by a pitcher. To see if this could be accepted for the model, similarly to Def.Tot from the hitter model, I checked the correlation between each of these variables and AAV. The correlation between BB9.1 and AAV is -0.34 and the correlation between HR9.1 and AAV is -0.13. Each of these variables do correlate negatively with AAV, meaning that their coefficients should be negative in a regression model.

Next, I checked each of these variables' correlation with the other predictors in the model. The positive coefficient for BB9.1 may have been due to multicollinearity, as it has a correlation coefficient of 0.63 with WHIP. However, the highest correlation coefficient that HR9.1 has with any other predictor is with WAR.1 and is not very large, at -0.32.

The incorrect signs of these two predictors makes the model difficult to interpret, so they should not be included. Removing either one of them did not cause the sign of the other to change, so I decided to remove both. The sum of squares due to these two variables, shown in the ANOVA table below, is pretty small. This means that after fitting the other predictor variables, these two predictors did not explain much of the variation in AAV.

After removing HR9.1 and BB9.1 from the model, the following coefficients summary and ANOVA table were obtained.

Table 4.5: Pitchers Model: Summary

Variable	Coefficient	t-value	$\Pr(>\|t\|)$
Intercept	13.87624	5.720	<0.0001
Age	-0.19146	-3.521	0.0006
WAR.1	2.24196	10.371	<0.0001
WAR.2	1.45225	7.806	<0.0001
WHIP.1	-2.96627	-2.865	0.0051
PosRP	-2.08069	-3.851	0.0002
PosSP	-0.24678	-0.372	0.7105

Table 4.6: Pitchers Model: Summary

Variable	DF	Sum Sq	Mean Sq	F-value	$\Pr(> F)$
WAR.1	1	2836.27	2836.27	754.8247	<0.0001
WAR.2	1	298.64	298.64	79.4775	<0.0001
Age	1	26.86	26.86	7.148	0.0087
WHIP.1	1	25.00	25.00	6.6525	0.0113
PosRP	1	72.89	72.89	19.3987	<0.0001
PosSP	1	0.52	0.52	0.1385	0.7105
Residuals	101	340.79	3.37		

All predictors are significant except for PosSP, which needs to be included for model interpretability if PosRP is included. I removed these two variables and fitted the model again and its CV error was 4.5965, compared to the CV error of 4.1464 with these two variables included, so I kept PosRP and PosSP. Other descriptive measures of this model are R^2=0.894, R^2_{adj}=0.889, AIC = 466.547, BIC = 488.151.

I checked the model's assumptions in the same manner as I did for the hitter model. The residuals appear to be randomly distributed around the horizontal line through the origin, and appear to be approximately normally distributed.

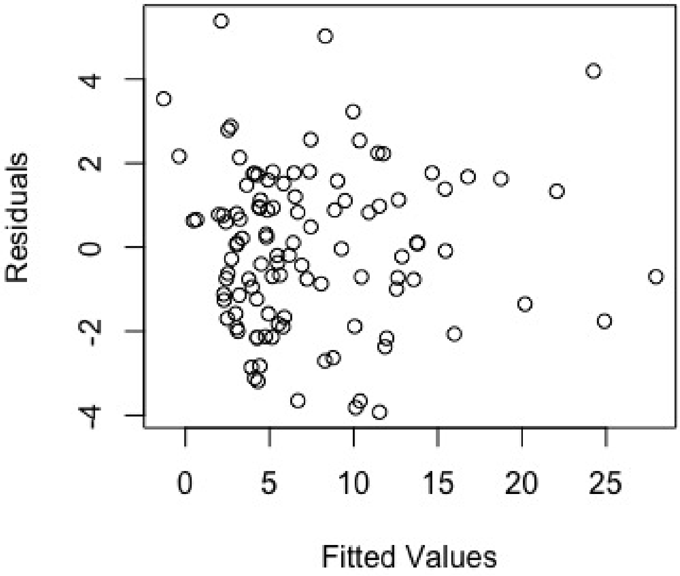

Figure 4.6: Pitchers Model: Residuals Plot

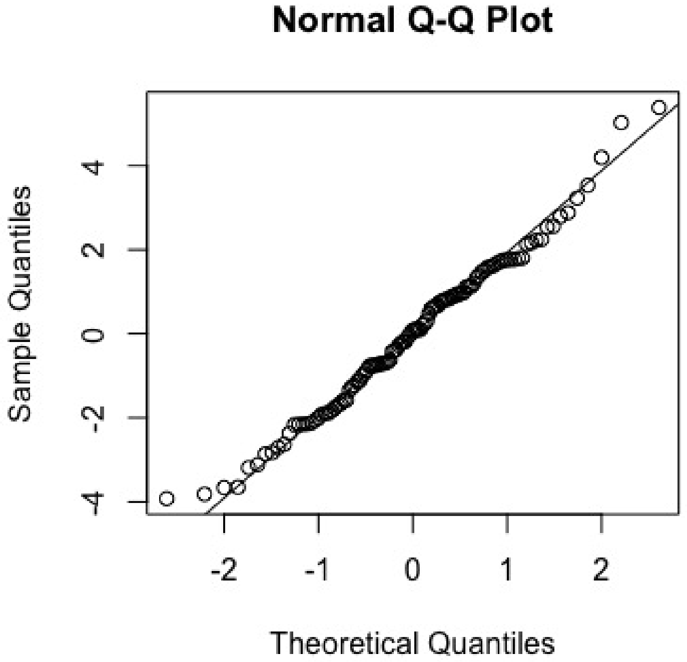

Figure 4.7: Pitchers Model: Normal Probability Plot

Since the model meets the assumptions, linear regression is appropriate for the fitted data, yielding the regression function

$$\begin{aligned}\text{AAV} =& 13.87624 - 0.19146 \times \text{Age} + 2.24196 \times \text{WAR.1} + 1.45225 \times \text{WAR.2} \\ & - 2.96627 \times \text{WHIP.1} - 2.08069 \times \text{PosRP} - 0.24678 \times \text{PosSP}.\end{aligned} \tag{4.3}$$

Chapter 5

Conclusion

5.1 Model Comparison

In the introduction of this report, I stated that my goal was to improve upon the results from (Magel and Hoffman, 2015). Their report contained one pitcher's model and one hitter's model using single-season performance metrics, and one pitcher's model and one hitter's model using career performance metrics. All four models predicted player salary. The single-season models were chosen for comparison since predicting salary using career data would not have any practical interpretability.

I fitted their models using my dataset for comparison. The comparison of the single-season hitter model performance is shown in the table below.

Table 5.1: Hitters Model Comparison

Criteria	My Model	M & H Model
CV Error	9.609	14.889
R^2_{adj}	.712	.47

It is also worth noting that their hitters model included 15 total variables, 8 of which had a p-value greater than .05.

The comparison of the single-season pitcher model performance is shown in the table below.

Table 5.2: Application

Criteria	My Model	M & H Model
CV Error	4.597	21.515
R^2_{adj}	.889	.481

Their pitchers model included 10 total variables, 5 of which had a p-value greater than .05.

For both models, I met my goal in creating models with more predictive power. Each of my models are also better fits to the fitted data, while being statistically sound, as shown by their compliance with the model assumptions.

5.2 Uses for the Models

In subsection (3.1.2) of this paper, I described the three main purposes of regression analysis. Here, I will discuss how my model fulfills these purposes.

The first main purpose, and my primary motivation for this project, was description. The presence of these models provides MLB fans, teams, agents, and any other parties involved with the sport with the knowledge of the specific performance metrics that have best predicted a player's AAV. This helps to bring the fans closer to understanding the thinking of MLB teams.

Not only are WAR.1 and WAR.2 very significant in both of my models, but they individually have high sums of squares values, meaning that they explain a lot of the variation in AAV. WAR is a metric that is commonly used by those who subscribe to sabermetrics, and we see that it is the most powerful predictor.

In subsection (3.1.3) of this paper, I discussed causality and regression analysis. Now that my models have been fit, it is important to not misinterpret their predictors as causes of AAV. The predictors have strong statistical relations with AAV, meaning that they contain much of the information that determines AAV. I consider the Age variable from the pitcher's model as an example, since it is simple to understand. The pitcher model states that a one year increase in a player's age leads to an average decrease in AAV of -0.19146 when holding all other predictors constant. This does not mean that the average MLB team intentionally offers a given pitcher $191,460 less than they would have, had the pitcher been one year younger. Although it is normal to be wary of aging players as they are closer to a decline in their physical abilities, age is likely correlated to other variables which would also be negatively correlated with AAV. An example is injury frequency. Older players are more likely to get injured and therefore get paid less. The model teaches us that a pitcher's age contains information that teams consider when determining how much a player is worth, but not necessarily that a pitcher's age causes teams to pay him less.

Although WAR is included in both models, the same precaution must be taken when interpreting teams' use of sabermetrics. We do not know if MLB teams are explicitly using WAR in valuing players, or if WAR contains information that teams are using. WAR is a very involved performance metric that is far more empirically involved than any traditional metric. The hitter version has batting, baserunning, and fielding components as well as positional and league adjustments, and the pitcher version has a component for pitching and for game context, as well as a league adjustment. These components and adjustments were obtained through previous statistical research, some of which included technology. At the very least, I think it is safe to say whether it be WAR or other metrics, MLB teams tend to consider advanced methodology in player evaluation whose complexity surpasses those of the traditional metrics.

The second main purpose was control. Although this was not one of my motivations in choosing this project topic, there are some applications. One of the most significant breakthroughs in sabermetrics was the value of a base on balls on offense. This had traditionally been viewed as a pitcher's error, rather than a hitter's achievement, but this seems to have mostly changed by now. Due to the evolving nature of analytics surrounding player performance evaluation, it is useful for both current and aspiring professional baseball players and their coaches to be aware of what MLB teams are looking for in a player. While caution is needed to not make a causality assumption, this model could be useful.

The third main purpose was prediction. This is useful for MLB front offices to project budgets and strategies. The former is quite simple, as MLB teams would be able to use this model to predict how much money they would have to spend on players that they want to sign in free agency, or if they will be able to afford to retain a player who is entering free agency. Teams could also use this for strategies in discovering market inefficiencies. If a team is confident that their own analytical tools are more accurate in measuring player performance, they could use this model to identify overvalued and undervalued players. This is essentially how the Oakland Athletics qualified for the playoffs in 2002 with the 3rd lowest payroll in the MLB.

Caution is required when making predictions with this model, since the data used is of contracts signed from 2015-2017. Player evaluation techniques are bound to continue to evolve with time, meaning that teams will alter their criteria for valuing players.

5.3 Future Work

If I was to improve my model in the future with more time and resources, some factors that I would consider follow.

I would attempt to include contract options, no-trade clauses, and other forms of non-guaranteed money in my contract valuation. Although contracts can contain non-monetary aspects, such as no-trade clauses, they make the signing more costly to the team.

I would expand this study to include contracts that were signed outside of free agency. This includes contract extensions and tendered contracts that are described in section (3.3).

Teams sometimes sign younger players to multi-year contract extensions prior to their eligibility for free agency that pay less than the player could have earned had he signed his contracts on a year-by-year basis. Players agree to this because it provides security in the event that they do not fulfill their potential or sustain a career-threatening injury. I would investigate this trade-off between security and maximizing salary, and attempt to predict AAV for such players.

Bibliography

[MH15] Rhonda Magel and Michael Hoffman. Predicting salaries of major league baseball players. *International Journal of Sports Science*, 5(2):51–58, 2015.

[NKNW96] John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

[Was13] Tyler Wasserman. Determinants of major league baseball player salaries. 2013.