ANALYSIS OF SALARY FOR MAJOR LEAGUE BASEBALL PLAYERS

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Michael Glenn Hoffman

In Partial Fulfillment
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

April 2014

Fargo, North Dakota

North Dakota State University
Graduate School

Title
AN ANALYSIS OF SALARY FOR

MAJOR LEAGUE BASEBALL PLAYERS

By

Michael Glenn Hoffman

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Rhonda Magel
Chair

Ronald Degges

Dragan Miljkovic

Approved:

4/7/2014	Rhonda Magel
Date	Department Chair

ABSTRACT

This thesis examines the salary of Major League Baseball (MLB) players and whether players are paid based on their on-the-field performance. Each salary was examined on both the yearly production and the overall career production of the player. Several different production statistics were collected for the 2010-2012 MLB seasons. A random sample of players was selected from each season and separate models were created for position players and pitchers. Significant production statistics that were helpful in predicting salary were selected for each different model. These models were deemed to be good models having a predictive r-squared value of at least 0.70 for each of the different models. After the regression models were found, the models were tested for accuracy by predicting the salaries of a random sample of players from the 2013 MLB season.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Rhonda Magel for her advice and guidance throughout the development of this thesis. Her knowledge and guidance were a valuable tool in the development of this thesis. I would also like to thank my other committee members, Dr. Ronald Degges, and Dr. Dragan Miljkovic for the time they took to review and provide feedback to this thesis. Additionally, I would like to thank all the other statistics professors I have had over the years to give me a base knowledge necessary to complete this research.

Finally, I would like to thank my family and my friends for their love and support over the years. My family has contributed to my success over the years and they continually reminded me the things that can happen if I push myself to my full potential. I would also like to thank my friends in the statistics department for their guidance, help and motivation over the course of this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDIX TABLES

CHAPTER 1. INTRODUCTION

As of 2013, Major League Baseball (MLB) has 30 clubs divided over 2 leagues and 3 divisions per league. In 2013, the baseball revenue for MLB was over $8 Billion with a wide variety of different markets. The markets range from very large markets of Los Angeles and New York to small markets like Pittsburgh and Milwaukee. ("*Mlb team values*," 2014). This revenue comes from a variety of sources but the two main drivers of each team's revenue is ticket sales and television contracts. The 2013 revenue information for each team is located in Appendix A and it is easy to see the wide difference of revenue from $167 Million to $471 Million. Due to the wide range of revenue that a team is able to collect based on their market, MLB and the Major League Baseball Players Association (MLBPA) agreed to a Revenue Sharing Plan in 2002. The MLB Revenue Sharing Plan involves each team paying 34% of Net Local Revenue into a 'pool of money'. Then the pool of money is evenly distributed between all 30 teams. Revenue Sharing was put into place to reduce some of the dominance of teams in large markets due to larger revenues and being able to attract all the top players due to higher revenue. This will be important in our analysis as it evens the payroll differences between teams and allows more equality between salaries of players and their production among all teams. ("*Basic agreement*." 2012).

Even though a team is able to collect a large amount of revenue, MLB wants to try and discourage teams from 'buying all the top players' to win a championship. Unlike other sports like the National Football League, MLB does not have a hard salary cap. Teams are allowed to spend as much as they please on salaries. However, Major League Baseball tries to discourage overspending by the enforcement of the Competitive Balance Tax which is often referred to as the Luxury Tax. The Competitive Balance Tax states that teams with a payroll over the tax threshold have to pay a penalty. In 2014, the threshold was equal to $189 Million. Teams with payrolls above the threshold will pay a penalty based on the number of consecutive years above the threshold. Current penalties are 17.5% for first year, 30% for 2 consecutive years, 40% for 3 consecutive years and 50% for 4 or more consecutive years above the tax threshold ("*Basic agreement,"* 2012). It is essential in our analysis that we assume that players are in an open market and could get the same money throughout the league so it is important to note that only the New York Yankees have paid the luxury tax every year and only 5 teams have ever paid a Competitive Balance Tax since it was implemented in 2003. (Axisa, 2013)

In Major League Baseball, there is the restriction of a minimum salary just like there is in the United States workforce. The minimum salaries for years 2010-2014 are listed in Appendix A, but range from $400,000 to $500,000 based on the calendar year. Minimum salary restrictions will not be a big factor in our analysis during rookies and players with less than 400 at-bats or 30 games pitched were excluded from our study. It is also worth noting that the maximum player salary for 2013 was $28,000,000 so there is a very wide range in salaries among players. Our goal is to describe the large differences in pay among players and if production on the field determines average yearly salary for a player. (Brown, 2012).

Players are also allowed to have an arbitration hearing for their salary if they have between 3 and 6 years of playing time in the MLB. This is necessary for these players because players cannot become free agents or switch teams in this time period. This should not affect our model too much because it should be assumed that players will be able to get a competitive wage regardless of what team they play for due to the revenue sharing agreement and an open market. (Axisa, 2013)

The last idea that needs to be introduced is difference in wage among different positions. As previously mentioned, MLB has two leagues with the main difference between the two leagues being the fact the pitcher does not bat in the American League (AL) and does bat in the National League (NL). In the American League, there is a Designated Hitter (DH) that does not play in the field but bats for the pitcher. In both leagues, there are 9 fielders on the field at a time. These positions are the Pitcher, First Baseman, Second Baseman, Shortstop, Third Baseman, Catcher and 3 Outfielders. The skill set that teams look for in a player varies by position. For example, first basemen on average are more of the power hitters on the team with slightly lower batting averages but a higher number of runs batted in and a higher number of home runs. On the other hand, second basemen are usually expected to have a higher batting average but a lower number of runs batted in and a lower home run production. Second basemen are normally faster and quicker players than first basemen. Therefore, it can be expected that the pay is different by position because a good first baseman might be worth more for a team than a good second baseman. These differences for each position will be accounted for in our model by an indicator variable for each position. The same thing will occur for pitchers as there are starting pitchers and relief pitchers. Therefore, it is good to have good starting pitchers and relief pitchers, but starting pitchers might be worth more. A good starting pitcher would pitch 6 or more innings in a game while a relief pitcher is needed to hold the lead for an inning or two in a game (Gelb, 2012).

In the rest of this paper, we will introduce some similar studies which have been conducted in Chapter 2. In Chapter 3, we will describe the details of the present study. In Chapter 4, we give the models that were developed based on the samples collected as described in Chapter 4. After deriving these models, we will use the models to predict salaries on a sample of 2013 players. The results of our predictions are given in Chapter 5. The overall conclusions are given in Chapter 6.

CHAPTER 2. LITERATURE REVIEW

There has an increasing amount of research on sports analytics and salary in particular. However, most of the research on salary has been conducted through analyzing the length of a player's contract and not on average yearly salary for a MLB player.

Moneyball (2011), a popular movie released in 2011 demonstrates the hard work that went into creating a roster on a limited budget for Billy Beane and the Oakland Athletics. In this movie, Billy Beane hires Peter Brand, a computer/statistical whiz, to use statistics to get the necessary production to be competitive and reach the playoffs while signing the players at a huge bargain. This movie is based on a true story of the 2002 Oakland Athletics where the Athletics were able to win 103 regular season games while only spending $39,679,746 which was the 3rd lowest payroll in the MLB in 2002. This was exactly the same number of wins as the New York Yankees in 2002, in which they spent $125,928,583 which was the highest payroll in MLB in 2002 (espn.com). Even though this was possible and a few teams try to get bargain players on a yearly basis, most teams in MLB have at least one player with a large contract indicating that if a team has a good player they will try to keep them at almost any cost to keep the fans interested in the team. (Orinick, 2014)

Meltzer (2005) conducted one of the few studies looking at the average yearly salaries for position players only. In this study, Meltzer conducts a 2-stage least-squares regression by running two regression models predicting the average yearly salary and the length of a contract. In these regression models, he uses the same independent variables for average yearly salary and length of contract which included 4 various production statistics and other variables such as all-star appearances, gold glove winnings, health status, age, position, contract status and payroll for the team in which the player play for that season. After finding these results for the first-stage regression models, he uses a two-stage regression predicting length of contract by average yearly salary which was found in first-stage regression model. Then he also predicted the average salary based on contract length and a subset of independent variables used in first regression. The study found that "performance metrics are a significant predictor of player salary" even though they only used On-Base Plus Slugging (OPS), and Plate Appearances as performance statistics in their study. In this paper, performance statistics will be the core of the study since the variables of health status and all-star appearances cannot be predicted accurately for future seasons when a contract is signed. Since it is believed that several performance statistics might be beneficial in predicting salary, this paper will focus on several performance statistics.

Turner and Hakes (2007) also conducted a study on average yearly salary. In their study, they again examined productivity and experience in determining salary for position players. They considered two main production determinants, Age and On-base plus Slugging (OPS) along with several extraneous variables such as MVP awards, ALLSTAR appearances and Population of Market along with several indicator variables for levels of negotiating freedom. They found that salary peaked at least 1.8 years after hitting productivity in baseball and salary declined slightly before retirement. It was also found that premier players had their performance decline slower than non-premier players. It is important in our study to look to see if these players near retirement are extreme cases in our model and if a salary of one of these players is different than the model would indicate. In our study, we accounted for the decrease in salary at the end of career by including many squared terms of the performance statistics. The model for pitchers was treated similarly and this model included many squared terms of performance pitching statistics.

Stankiewicz (2009) looked into the length of a contract and the productivity of a player. In this paper, a weighted offensive statistic was used and regressed over age, games played and coaching success. The weighted offensive statistic used in her paper is called Equivalent Average (EqA) and calculated as the following:

$$EqA = \frac{hits + total\ bases + 1.5 * (walks + hits\ by\ pitch) + stolen\ bases}{(total\ at\ bats) + walks + hit\ by\ pitch + number\ caught\ stealing + \frac{stolen\ bases}{3}}$$

This Equivalent Average statistic was not created by Stankiewicz but by Clay Davenport and explained in her paper. This statistic is actually similar to OPS except that this statistic takes into account stolen bases whereas OPS is purely a power statistic. In this paper, OPS will be used because it is commonly known and used in many papers whereas EqA is not commonly used. The main goal of Stankiewicz's paper was to determine if players with long term contracts were actually more productive than players with one year contracts. It was found that players with long contracts (greater than 1 year) were in fact more productive and had a higher EqA. The drawback from this study is that salaries of players were not examined. Therefore, Stankiewicz's research (2009) lacked the relationship between production and salary whether or not players are overpaid based on their production which will be the focus of this paper.

One final important study that was conducted in 2011, examined that effect of contract year performance and the future production of the player and salary that the player received. Hochberg (2011) used data from 1993-2010 to examine the salary of position players. He used the statistics of: on-base plus slugging (OPS), stolen base rate, fielding rate above replacement, age and position, using a linear model and a linear model with interaction. In order to capture the effect of production over the last few years Hochberg used the value for each player's OPS over the previous 3 seasons. It was found in his study that position did not influence the salary. Hochberg also found that “performance deviates from year to year, the more susceptible individuals will be to underweight past performance data”. This indicates that contract year performance is overvalued compared to the previous seasons before a contract was signed. This paper will not look at contract year performance but examine salary based on yearly performance and career performance.

CHAPTER 3. DESIGN OF STUDY

The goal of this study is to determine the significant statistics that determine the average yearly salary of a MLB player. In this study, pitchers and position players are judged off total different statistics in a game. Separate models for pitchers and position players will be derived using least squares regression with the stepwise selection technique. Indicator variables will be included in the models to indicate the positions of the players. If a player played more than one position in a season, the position with highest amount of innings played will be used.

Data was collected from three MLB seasons which included the 2010, 2011 and 2012 seasons. Players with limited time in the major leagues were omitted. A player was omitted from the position player model if they had fewer than 400 at-bats. A player was omitted from the pitching model if they pitched fewer than 30 games. A restriction for the games pitched instead of innings pitched was used because innings per performance would not be equal for relief pitchers and starting pitchers. These thresholds were used because these appearances accounted for approximately 1 full year of playing time for a full-time player. Since it was believed that salary might vary between the two leagues a stratified random sample was used with 90 players selected from each league for both position players and for pitchers. This resulted in approximately 5-6 players on average from each team selected. The resulting sample size was 540 players to develop the regression model for position players and 540 pitchers to develop the regression model for pitchers.

The data was collected from 3 websites. Salaries were collected from two different websites because some salaries were not listed on one website but on the other website. These two websites included Baseball Player Salaries (baseballplayersalaries.com) and Baseball Reference (baseball-reference.com). All the production statistics were then collected from

Baseball Reference (baseball-reference.com). Finally, some of the differentiations between starting pitchers and relief pitchers were found on the ESPN website (espn.com).

The goal of this study is to examine the significant factors for all players in determining average yearly salary. Different statistics need to be evaluated for pitchers and position players. Pitchers cannot be evaluated on the same statistics especially since pitchers in the American League do not bat. Separate models would be created for pitchers and position players. To account for the differences in position, an indicator variable will be used for each position. Therefore, there will be 6 indicator variables introduced in the model for position players and one indicator variable for the model for pitchers, differentiating between starting pitchers and relief pitchers. There was not differentiation made between the three outfield positions since many players played more than one outfield position throughout the course of the season with little skill set differences between the outfielders. The role of the relief pitcher was not accounted for in this study and set-up pitchers and closers were grouped into the same position (relief pitchers).

There will be two models created for position players and two more models created for pitchers. The first set of models for each player will only use yearly statistics and try to predict the yearly salary of a given player for that season. The models using only yearly statistics are helpful in determining if a player is attaining production statistics based on their salary. These models are helpful but cannot be used for prediction since the yearly statistics a player receives are not known in advance. The second set of models that will be developed will include only career statistics that a player has accumulated in the past seasons. The models using career statistics should indicate whether players are getting salaries that are based on their career statistics. These models will be useful for evaluating salary and also useful in predicting the future salary of a given player.

In this study, there are many different statistics examined for production instead of only using only one production statistic such as OPS or EqA, as in previous studies. The statistics that were chosen were selected for two main reasons. These two reasons included that baseball fans were familiar with the statistics and the statistics were easily accessible at baseball-reference.com. In order to adjust for the different variances in salary for a given set of production statistics and to make a better predictive model, the log of salary will be used for the dependent variable in all of the models. All the hitting statistics and pitching statistics chosen for consideration are listed in Appendix B. Many different combinations of these variables along with stepwise selection with an entry and exit value of 0.15 will be used to determine the significant variables in each model. All squared variables that were considered were based on a diagnosis of a scatterplot of each predictor variable against log(salary). Only relationships in which there appeared to be a quadratic relationship were considered for a squared term. After the variables are selected and the models are developed, the best models will then be used to predict the salary of pitchers and position players for the 2013 season. Our predictions will include predictions from a stratified random sample of 90 position players from each league and 90 pitchers from each league in 2013.

CHAPTER 4. RESULTS

4.1. Yearly Position Player Model

The first model that was developed was a model for the salary of a player based on their performance statistics for that given year. This model will not be able to be used to predict the yearly salary as the yearly statistics for a player are not known before the season. However, this model is good for two reasons. First of all, the model could be used to see if players are performing in accordance to their salary, and secondly it could be used to determine what yearly performance statistics are good predictors of salary.

The stepwise selection technique with an entry and exit level of 0.15 was used to determine the significant yearly performance statistics. The significant yearly performance statistics include: Runs Batted In (RBI), Triples, Games, Games Squared, Plate Appearances, Sacrifice Hits, Strikeouts, Position, Doubles and Doubles Squared. It is interesting to note in this model that position of the player and the year (2010, 2011 or 2012) are not significant. This model does not appear to be explaining the variation in salary very well as only 35% of the variation in salary is explained by this yearly statistics model in Appendix A. Additionally, this finding would imply that players are not performing based on their salaries. This would mean that many players are being overpaid based on their production statistics while other players are being underpaid. From residual diagnosis, it is found that 261 players have a negative residual (underpaid) and 279 players have a positive residual (overpaid). Therefore, it appears that roughly an equal number of players are under performing and over performing based on their salary. This would also indicate that general managers are not able to accurately predict how much to pay players based on their future production. The model could also have low predictive power since it does not include variables of health (injury) or performance incentives such as

MVP awards or all-star appearances. Additionally, players may have an awesome year production wise or a year of struggles that could not be predicted by previous career performance. In baseball, it is common for even the premier players to have a bad year once in a while with players having a change of 10 HR or have their batting average drop or increase by 0.040 some years which would affect the salary prediction significantly.

4.2. Yearly Pitchers Model

After the model using yearly production statistics for the position players was developed, a similar model was developed using yearly production statistics for pitchers. The model was then used to estimate the salary for a pitcher on a given year if his yearly production were known. Similar to the model for position players, these models will not be used to predict salary as yearly performance statistics are not known in advance. However, yearly models will be helpful in determining if pitchers are performing close to their expectations.

The stepwise selection technique was used to determine the significant yearly production statistics in this model with an entry and exit level of 0.15. The significant yearly performance statistics include: Saves, Saves Squared, Year, Games Started, Walks Allowed, Walks Allowed Squared, Balks, Balks Squared, Losses, Complete Games and Complete Games Squared. This model was developed based on a sample size of 539 since one observation (James Shields in 2011) was deleted from the sample based on a large Cook's Distance (Abraham & Ledolter, 2006). James Shields has a salary that was very extreme compared to his performance in 2011 and was eliminated to avoid bias in the salary estimates. This model as a whole is even worse at explaining the salary of a pitcher than the position player's model with an R-squared value of only 0.2506. This model actually only explains roughly 25% of the variation in salary so there are many pitchers being underpaid and overpaid on a yearly basis. From a diagnosis of the residuals, it’s found that 263 have negative residuals (underpaid) and 276 pitchers have positive residuals (overpaid) from our model. Therefore, there doesn’t appear to be a difference in the number of players being overpaid or underpaid but a wide spread in performance and salary. This model also supports the opinion of many other authors who were reluctant to model pitcher’s salary because of the extreme differences in pitcher statistics due to their role in the game (Meltzer, 2005). There are many different types of pitchers: starting pitchers, long-relief pitchers, closers, short-relief pitchers, fill-in relievers and many others. This appears true with the yearly pitching model as the position variable (starting or relief pitcher) was not even selected to stay in the model and when it was put in the model it did not significantly help predict salary. A diagnosis of the model including career statistics for pitchers will need to be examined to determine if position/role of the pitcher is actually necessary for accurate predictions along with hopefully attaining a more powerful model in explaining salary.

4.3. Career Position Players Model

The yearly regression models did not appear to be very good for explaining the salary of a player based on their yearly production statistics. In addition, these models are not helpful in regards to trying to predict a player’s salary since the yearly production statistics of a player is unknown at the time of signing a contract. Prediction models could be created to predict these yearly statistics, but they would include career statistics so it would appear logical to simply predict a player’s salary based on a players career production statistics. In order to find career production statistics for each player, yearly statistics were aggregated from 1984 to the year before salary was examined. Therefore, if a player’s salary was examined in 2010, then career statistics were aggregated up to 2009 for that player. The same production statistics that were considered for the year were considered for the entire career for each player. The stepwise regression technique with an entry or exit level of 0.15 was used to develop the model based on career production statistics. The significant predictors are career production statistics for the following: Total Bases, Total Bases Squared, Games, Games Squared, Sacrifice Hits, Sacrifice Hits Squared, Position, Caught Stealing, Caught Stealing Squared, Runs, Runs Squared, Ground into Double Plays, Ground into Double Plays Squared, At-Bats, Stolen Bases, and Year. The regression output is found in Appendix A for this model. Immediately, it is found that year and position of the player are now significant predictors in salary. This would make sense that salary should increase at least by a little bit each year for inflation. It also would appear that some positions would demand higher salaries as indicated in this model since teams might be willing to pay a top first baseman different than a top second baseman or vice versa. Recall, that in the yearly model both year and position were insignificant in predicting salary.

When the regression model was developed, it was found that 4 players did not fit the model very well. These players included: Gary Matthews, Ivan Rodriguez and Ken Griffey Jr. from 2010 and Ivan Rodriguez from 2011. All of these players were discarded from the sample based on their Cook's Distance values. Eliminating these players was critical to remove any bias in our estimates. This regression model was then developed with a sample of 536 observations and yielded an R-Squared value of 0.7423. This model explained the variation in salary of players much better than the model using yearly production statistics which had an R-Squared value of 0.3543. The predicted R-squared value further validates the strong predictive power of this model with a predictive R-Squared value of 0.7155. This model will be used to predict the salaries of position players for a sample of 2013 players since career statistics are known before a contract is signed. The predictions using this model will be discussed in Chapter 5.

4.4. Career Pitchers Model

The model using yearly production statistics for pitchers is not satisfactory for predicting salaries of pitchers since these yearly production statistics are not known in advance. Therefore, a career statistics model for pitchers will be considered which will be similar to the career statistics model for position. Again, this was done by collecting yearly production statistics from 1984 to the year previous to when the player's salary is examined. The same production statistics were considered in the career statistics model as in the yearly statistics model. The stepwise regression technique with an entry or exit level of 0.15 was used. The significant predictors for this model included career statistics for the following categories: Strike Outs, Strike Outs Squared, Walks, Year, Wild Pitch, Wild Pitch Squared, Saves, Saves Squared, Games Finished, Games Finished Squared, Intentional Walks, Intentional Walks Squared, Dominance, Earned Run Average, Games, Games Squared, Pitcher Role, Home Runs Allowed, Losses, and Losses Squared. The regression output with regression estimates are found in Appendix A for this model.

When the regression model was run it was found that 2 players had especially high Cook's Distances and appeared to have heavy influences on the estimates of the coefficients. These players were Mariano Rivera in 2011 and Billy Wagner in 2010. Both of these players were in the last couple years of their career. Previous research warned about this problem and it was concluded that these players had salaries not in line with other players and their performance. These players could possibly have their salary based partly based on their fan popularity and All-Star appearances which was not considered in these models. Therefore, these players were excluded from the model to reduce any bias in regression coefficients. The regression model was then estimated with a sample of 538 observations and yielded an R-squared value of 0.7097. This model explained the salary of players much better than the model using yearly production statistics with an R-squared value of 0.2506. Unlike the model using year statistics, the role of the pitcher (starting or relief pitcher) was significant. Based on the assumption, that salary of pitchers is so varied that other researchers did not even consider pitcher salaries, it appears the model is good even though it has slightly lower predictive power than the position players model with a predictive R-squared value of 0.6872. The lower R-squared in this model compared to the position players model is mostly likely caused by the varying roles of the pitcher. This problem would never be fixed though, since the role of a pitcher often is switched during different periods in the season. A pitcher could be a long-relief pitcher for part of the season and set-up relief pitcher later in the season. Therefore, it would be hard to control for so many roles of a pitcher. Since the R-squared and predicted R-squared value are fairly close to each other, this model should do a good job at predicting salaries in future seasons. This model will be used to predict the salaries of a pitcher for a sample of pitchers from the 2013 season.

CHAPTER 5. PREDICTION

In Chapter 4, it was found that the best model for being able to predict a player's salary included career production statistics that the player had accumulated up to that point in their career. These models were found in sections 4.3 and 4.4. These models will now be used to test how well they can predict the salaries of a stratified sample of players from 2013 based on their career statistics.

To test the models a stratified random sample of 90 players from each league was sampled for both the model for the pitchers and for the model for the position players. This random sample yielded 180 players for each model to be used to predict the salaries. Two stratified random samples included players from all 30 MLB teams in 2013. These datasets yielded players from each position on the field but did not yield any players in the designated hitter role. This occurred by random chance that no designated hitters players were selected and is likely due to the low numbers in players in designated hitter role. The number of players for each position selected in the random sample for both the pitchers model and the position player's model is found in Appendix A.

5.1. Predictions for Pitchers

The prediction for a pitcher's salary is based on the career statistics of the player before 2013. The original regression model was based on the natural log of salary to adjust the model due to unequal variances. Our predictions will be on the natural log scale also. Since this is a linear regression model, an example prediction is easy to show. C.C. Sabathia is a fairly well-known pitcher. In Table 5.1, the regression parameters and C.C. Sabathia's career statistics before 2013 are displayed:

Table 5.1. Example C.C. Sabathia

Model	Regression Parameters	C.C.'s Career Stats
Intercept	13.7477	1
SOCAR	0.005442255	2235
SOCAR2	-0.000001484	4995225
BBCAR	-0.002013706	767
Year	0.19384	3
WPCAR2	-0.000284407	2500
SVCAR	0.018909	0
GFCAR2	-0.00000506	0
GFCAR	-0.005148328	0
IBBCAR2	-0.000401642	961
IBBCAR	0.009380998	31
DOMINANCECAR	-0.10565	0.86722
ERACAR	-0.095431	4.7773
WPCAR	0.01984	50
SVCAR2	-0.000026457	0
GCAR	0.003608008	385
GCAR2	-0.000003292	148225
POSSTART	0.22913	1
HRCAR	-0.004434099	229
LCAR2	0.000103331	10816
LCAR	-0.013161	104

In order to find C.C. Sabathia’s predicted salary for 2013, you would simply multiply each estimated parameter by the associated career statistic and add these values up. The regression model in this case would be:

ln(salary) =13.7477+0.005442255*SOCAR-0.000001484*SOCAR2-0.002013706*BBCAR+ 0.19384*YEAR-0.000284407*WPCAR2+0.018909*SVCAR-0.00000506*GFCAR2-0.005148328*GFCAR-.000401642*IBBCAR2+ 0.009380998*IBBCAR-0.10565*DOMINANCECAR -0.095431*ERACAR+ 0.01984*WPCAR-0.000026457*SVCAR2+0.003608008*GCAR-

0.000003292*GCAR2+ 0.22913*POSSTART -0.004434099*HRCAR+

0.000103331*LCAR2-0.013161*LCAR

Substituting C.C. Sabathia's career statistics for each of the different categories into the regression equation, the following is obtained:

ln(salary) =13.7477+0.005442255*2235-0.000001484*4995225-

0.002013706*767+0.19384*3-0.000284407* 2500+0.018909*0-0.00000506*0-

0.005148328*0-0.000401642*961+0.009380998*31-0.10565*0.86722 -

0.095431*4.7773+0.01984*50-0.000026457*0+0.003608008*385-

0.000003292*148225+0.22913*1 -0.004434099*229+0.000103331*10816-

0.013161*104=17.03726

Since there is such a wide range of salaries, it would be good to see how far off our predictions are from the actual salaries in terms of percent error in terms of ln(salary) since that is how our predictions were made. The following is the calculated percent error for C.C. Sabathia:

Percent Error=$\frac{|estimated\ salary-actual\ salary|}{actual\ salary}=\frac{|17.03726-16.951|}{16.951}$=0.51%

The predicted salary for C.C. Sabathia in 2013 is 17.037 on the natural log scale. However, the salary for C.C. Sabathia in 2013 is actually 16.951 on the natural log scale. This would tell us that our model states that C.C. Sabathia is actually underpaid for his performance on the field. Looking at the percent error for the salary of C.C. Sabathia , it is found that the salary prediction is actually 0.51% above the actual salary for C.C. Sabathia. The accuracy for all 180 pitchers in 2013 sample is found below in Table 5.2 with most of the predictions within 9% of the true salary for players on the log salary scale.

Table 5.2. Accuracy of Pitcher Predictions

Percent Error	# of Players
0-3%	42
3-6%	39
6-9%	34
9-12%	37
12-15%	20
15+%	8

5.2. Predictions for Position Players

The salary predictions for position players are based on the career statistics of the player before 2013. The original regression model was based on the log of the salary to adjust for unequal variances so the predictions will be on the natural log scale. Since this is a linear regression model, an example prediction is easy to show. Joe Mauer is a fairly well-known position player. In Table 5.3, the estimated regression parameters and Joe Mauer's career statistics before 2013 are listed. In order to find his predicted salary for 2013, you would simply multiply the estimated regression parameter by the corresponding career statistic and add them up. The regression model in this case would be:

$$\begin{aligned}&\ln(\mathrm{salary})= 12.4865+0.004355*\mathrm{TBCAR}-0.000001418*\mathrm{GCAR2}+0.000888548*\mathrm{GCAR}-\\&0.000000468*\mathrm{TBCAR2}+0.000086283*\mathrm{SHCAR2}+0.000106329*\mathrm{SHCAR}-\\&0.32618*\mathrm{POS1B}+0.13093*\mathrm{POS2B} + 0.10874*\mathrm{POSSS}+0.082688*\mathrm{POS3B}-\\&0.000201974*\mathrm{POSDH}+0.30243*\mathrm{POSC}-0.00006925*\mathrm{CSCAR2} -\\&0.00404255*\mathrm{CSCAR}+0.000001771*\mathrm{RCAR2}-0.000018625*\mathrm{RCAR}+\\&0.000013946*\mathrm{GDPCAR2} +0.001353982*\mathrm{GDPCAR}-0.00082267*\mathrm{ABCAR}+\\&0.003986277*\mathrm{SBCAR}+0.061845*\mathrm{YEAR}\end{aligned}$$

Table 5.3. Example Joe Mauer

Model	Regression Parameters	Joe's Stats
Intercept	12.4865	
TBCAR	0.004354684	1839
GCAR2	-0.000001418	1134225
GCAR	0.000888548	1065
TBCAR2	-0.000000468	3381921
SHCAR2	0.000086283	16
SHCAR	0.000106329	4
POS1B	-0.32618	0
POS2B	0.13093	0
POSSS	0.10874	0
POS3B	0.082688	0
POSDH	-0.000201974	0
POSC	0.30243	1
CSCAR2	-0.00006925	225
CSCAR	-0.00404255	15
RCAR2	0.000001771	391876
RCAR	-0.000018625	626
GDPCAR2	0.000013946	16900
GDPCAR	0.001353982	130
ABCAR	-0.00082267	3933
SBCAR	0.003986277	43
year	0.061845	3

Substituting Joe Mauer's career statistics for each of the different categories into the regression equation, the following is obtained:

$$\ln(salary)= 12.4865+0.004355*1839-0.000001418*1134225+0.000888548*1065-0.000000468*3381921+0.000086283*16+0.000106329*4-0.32618*0+0.13093*0 + 0.10874*0+0.082688*0-0.000201974*0+0.30243*1-0.00006925*225-0.00404255*15+ 0.000001771*391876-0.000018625*626+0.000013946*16900 +0.001353982*130-0.00082267*3933+0.003986277*43+0.061845*3= 16.69346$$

Again since there is such a wide range of salaries, percent error will be used to evaluate the accuracy of predictions. The percent error of the natural log of salary for Joe Mauer is found below:

Percent Error=$\frac{|estimated\ salary - actual\ salary|}{actual\ salary} = \frac{|16.69346-16.951|}{16.951}$=1.52%

Therefore, the predicted salary for Joe Mauer in 2013 is 16.69346 on the natural log scale. However, the salary for Joe Mauer in 2013 is actually 16.951 on the natural log scale. This would tell us that our model states that Joe Mauer is overpaid for his performance on the field. When looking at the accuracy of the prediction, it is found that the percent error for the salary of Joe Mauer is found to be 1.52% above the actual salary for Joe Mauer on the natural log scale. The accuracy for all 180 position players in the 2013 sample for position players is found below in Table 5.4. From the table, it is found that with most of the predictions are within 3% of the actual salary on the natural log scale and all but 9 players have a salary prediction within 9% of their actual salary on this scale. These predictions are more accurate for position players than for pitchers. The predictions for both the pitchers and position players appear to be relatively good.

Table 5.4. Accuracy of Position Player Predictions

Percent Error	**# of Players**
0-3%	97
3-6%	57
6-9%	17
9-12%	3
12-15%	3
15+%	3

CHAPTER 6. CONCLUSIONS

After fitting models and predicting the salaries of random samples of position players and pitchers, it was found that the models using career production statistics were the most useful for both pitchers and position players since these statistics are known in advance of signing a player to a contract. Several career production statistics were included in both models. It was found that the significant career performance statistics for position players included: Total Bases, Total Bases Squared, Games, Games Squared, Sacrifice Hits, Sacrifice Hits Squared, Position, Caught Stealing, Caught Stealing Squared, Runs, Runs Squared, Ground into Double Play, Ground into Double Plays Squared, At-Bats, Stolen Bases and Year. A different subset of performance statistics was found to be significant in determining the salaries of pitchers. These career performance statistics included: Strike Outs, Strike Outs Squared, Walks, Year, Wild Pitch, Wild Pitch Squared, Saves, Saves Squared, Games Finished, Games Finished Squared, Intentional Walks, Intentional Walks Squared, Dominance, Earned Run Average, Games, Games Squared, Pitcher Role, Home Runs Allowed, Losses, and Losses Squared. The career performance statistics led to a high predictive power in determining salaries as the predictive R-squared values was 0.7155 and 0.6872 for the position players and pitchers respectively.

After developing the models to predict salaries for both pitchers and position players, predictions of salaries were made for a random sample of 180 pitchers and a random sample of 180 position players. It was found that prediction errors were within 0% to 12% for approximately 84% of the pitchers and approximately 96% for position players. Therefore, career production statistics appear to be large influence on the salary for players. It was also found that players in the last few years of their career did not fit the model very well since these players were paid either lower than their career statistics would indicate or paid higher than their career statistics would suggest due to fan popularity and mentorship they could offer to younger players on the team. Examples of these players included Mariano Rivera and Billy Wagner.

There are many areas of future research that could be done. Future research could examine health status variables and number and type of awards won. It might be beneficial to look into All-Star game appearances, Most Valuable Player (MVP) awards, Gold Glove awards, and possibly the number of days on the disabled list in their career. Award winnings will likely increase the salary of these players significantly but also only apply to a relatively small portion of players. Research could examine the effect of performance bonuses for players. Many players today sign a contract with incentives to produce certain performance. It would be interesting to see if these contracts significantly increase the production statistics for these players or if it is better not to have contracts of this type.

Two other areas of future research include contract length and arbitration player salary. Previously contract length was modeled using a 2-stage regression with salary. It would be interesting to see if contract length is based on career statistics and age or if agents are able to persuade teams to offer a significantly longer contract regardless of past performance. Additionally, it would be interesting to see if arbitration significantly reduces pay for young players or if court rulings give players in arbitration roughly the market value salary.

For any model, there is a wide range of factors that each team examines when offering a player a contract. A team will consider past performance, expected future performance, contract length, and fan reaction to the signing of a player. In this paper, both yearly performance statistics and career performance statistics were examined and it was found that salary was significantly influenced by a variety of career performance statistics and not as much by yearly performance statistics.

REFERENCES

Abraham, B., & Ledolter, J. (2006). *Introduction to regression modeling*. Belmont, California: Thomson Brooks/Cole.

Axisa, M. (2013, 09 11). *Report: Yankees set to pay record $29.1 million luxury tax bill*. Retrieved from http://www.cbssports.com/mlb/eye-on-baseball/23592770/report-yankees-set-to-pay-record-291-million-luxury-tax-bill

Basic agreement. (2012). Retrieved from http://mlb.mlb.com/pa/pdf/cba_english.pdf

Baseball player salaries. (2014). Retrieved from http://baseballplayersalaries.com/

Baseball-reference. (2014). Retrieved from baseball-reference.com

Brown, M. (2012, June 04). *Bizball*. Retrieved from http://www.baseballprospectus.com/article.php?articleid=17225

Espn. (2014). Retrieved from espn.com

Hakes, J.K. & Turner, C. (2011) Pay, productivity and aging in major league baseball. *Journal of Productivity Analysis*. 35, 61 - 74.

Hochberg, Daniel. "The Effect of Contract Year Performance on Free Agent Salary in Major League Baseball" 2011, Department of Economics, Haverford College, PA

Gelb, M. (2012, 09 11). *Understanding payroll and luxury tax*. Retrieved from http://www.philly.com/philly/blogs/phillies_zone/understanding-payroll-and-luxury-tax.html

Meltzer, Josh. "Average Salary and Contract Length in Major League Baseball: When Do They Diverge?" 2005, Department of Economics, Stanford University, CA.

Moneyball [DVD]. (2011).

Mlb salaries. (2014). Retrieved from http://www.cbssports.com/mlb/salaries

Mlb team values. (2014). Retrieved from http://www.forbes.com/mlb-valuations/list/

Orinick, S. (2014). *Mlb team payrolls*. Retrieved from http://www.stevetheump.com/Payrolls.htm

Stankiewicz, Katie (2009) "Length of Contracts and the Effect on the Performance of MLB Players," *The Park Place Economist*: Vol. 17

APPENDIX A. REVENUE AND MODEL OUTPUT

Table A1. 2013 Revenue by Team

TEAM	REVENUE	PAYROLL
NEW YORK YANKEES	$471,000,000	$228,835,490
BOSTON RED SOX	$336,000,000	$150,655,500
PHILDEPHIA PHILLIES	$279,000,000	$165,385,714
CHICAGO CUBS	$274,000,000	$104,304,676
SAN FRANSICSO GIANTS	$262,000,000	$140,264,334
LOS ANGELES DODGERS	$245,000,000	$216,597,577
ST LOUIS CARDINALS	$239,000,000	$115,222,086
LOS ANGELES ANGELS OF ANAHEIM	$239,000,000	$127,896,250
TEXAS RANGERS	$239,000,000	$114,090,100
DETROIT TIGERS	$238,000,000	$148,414,500
NEW YORK METS	$232,000,000	$73,396,649
ATLANTA BRAVES	$225,000,000	$89,778,192
WASHINGTON NATIONALS	$225,000,000	$114,056,769
CHICAGO WHITE SOX	$216,000,000	$119,073,277
SEATTLE MARINERS	$215,000,000	$72,031,143
MINNESOTA TWINS	$214,000,000	$75,802,500
BALTIMORE ORIOLES	$206,000,000	$90,993,333
TORONTO BLUE JAYS	$203,000,000	$117,527,800
CINCINNATI REDS	$202,000,000	$107,491,305
MILWAUKEE BREWERS	$201,000,000	$82,976,944
COLORADO ROCKIES	$199,000,000	$71,924,071
HOUSTON ASTROS	$196,000,000	$22,062,600
MIAMI MARLINS	$195,000,000	$36,341,900
ARIZONA DIAMONDBACKS	$195,000,000	$89,100,500
SAN DIEGO PADRES	$189,000,000	$67,143,600
CLEVELAND INDIANS	$186,000,000	$77,772,800
PITTSBURGH PIRATES	$178,000,000	$79,555,000
OAKLAND ATHLETICS	$173,000,000	$60,664,500
KANAS CITY ROYALS	$169,000,000	$81,491,725
TAMPA BAY RAYS	$167,000,000	$57,895,272

Table A2. Minimum Salary by Year

YEAR	MINIMUM SALARY
2010	$400,000
2011	$414,000
2012	$480,000
2013	$490,000
2014	$500,000

Table A3. Number of Position Players by Position in 2010-2012 Dataset

Position	Number of Players
First Base (1B)	57
Second Base (2B)	67
Shortstop (SS)	60
Third Base (3B)	66
Outfield (OF)	208
Catcher (C)	78
Designated Hitter (DH)	4

Table A4. Number Pitchers by Role in 2010-2012 Dataset

Position	Number of Pitchers
Starting Pitcher (SP)	210
Relief Pitcher (RP)	330

Table A5. Number of Position Players by Position in 2013 Dataset

Position	Number of Players
First Base (1B)	25
Second Base (2B)	19
Third Base (3B)	23
Shortstop (SS)	16
Outfield (OF)	63
Catcher (C)	34

Table A6. Number Pitchers by Role in 2013 Dataset

Position	Number of Pitchers
Starting Pitcher (SP)	71
Relief Pitcher (RP)	109

Table A7. Yearly Position Players Model Output

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	13.89073	0.18661	74.44	<.0001
RBI	0.01559	0.00385	4.05	<.0001
TR	-0.08882	0.02134	-4.16	<.0001
G2	-0.00012807	0.00003198	-4.00	<.0001
G	-0.00445	0.00570	-0.78	0.4344
PA	0.00911	0.00110	8.27	<.0001
SH	-0.07106	0.02019	-3.52	0.0005
SO	-0.00544	0.00191	-2.85	0.0046
POSC	-0.33131	0.13937	-2.38	0.0178
POS1B	-0.21493	0.15497	-1.39	0.1661
POS2B	-0.11803	0.14220	-0.83	0.4069
POS3B	-0.10895	0.14439	-0.75	0.4509
POSSS	-0.09487	0.15618	-0.61	0.5438
POSDH	0.44384	0.50299	0.88	0.3780
DB	-0.04589	0.02371	-1.94	0.0534
DB2	0.00038281	0.00041023	0.93	0.3512

Table A8. Career Position Players Model Output

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	12.48645	0.09888	126.27	<.0001
TBCAR	0.00435	0.00064295	6.77	<.0001
GCAR2	-0.00000142	2.24958E-7	-6.30	<.0001
GCAR	0.00088855	0.00049242	1.80	0.0717
TBCAR2	-4.67613E-7	9.765175E-8	-4.79	<.0001
SHCAR2	0.00008628	0.00001779	4.85	<.0001
SHCAR	0.00010633	0.00347	0.03	0.9756
POS1B	-0.32618	0.10743	-3.04	0.0025
POS2B	0.13093	0.09619	1.36	0.1740
POSSS	0.10874	0.10561	1.03	0.3036
POS3B	0.08269	0.09498	0.87	0.3844
POSDH	-0.00020197	0.33388	-0.00	0.9995
POSC	0.30243	0.09883	3.06	0.0023
CSCAR2	-0.00006925	0.00002219	-3.12	0.0019
CSCAR	-0.00404	0.00473	-0.85	0.3936
RCAR2	0.00000177	6.461395E-7	2.74	0.0063
RCAR	-0.00001862	0.00135	-0.01	0.9890
GDPCAR2	0.00001395	0.00001327	1.05	0.2938
GDPCAR	0.00135	0.00356	0.38	0.7038
ABCAR	-0.00082267	0.00018418	-4.47	<.0001
SBCAR	0.00399	0.00117	3.39	0.0007
Year	0.06184	0.03368	1.84	0.0669

Table A9. Yearly Pitchers Model Output

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	13.71151	0.09417	145.61	<.0001
SV	0.06208	0.01684	3.69	0.0003
YEAR	0.19649	0.05464	3.60	0.0004
GS	0.04191	0.00852	4.92	<.0001
BB2	-0.00011802	0.00004017	-2.94	0.0034
SV2	-0.00085850	0.00042813	-2.01	0.0455
BK2	-0.03495	0.04973	-0.70	0.4826
CG2	0.02585	0.01754	1.47	0.1411
L	0.03585	0.02085	1.72	0.0861
BK	-0.05048	0.14118	-0.36	0.7208
CG	-0.04096	0.11007	-0.37	0.7099

Table A10. Career Pitchers Model Output

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	13.74771	0.30084	45.70	<.0001
SOCAR	0.00544	0.00066403	8.20	<.0001
SOCAR2	-0.00000148	2.288632E-7	-6.48	<.0001
BBCAR	-0.00201	0.00058018	-3.47	0.0006
YEAR	0.19384	0.03469	5.59	<.0001
WPCAR2	-0.00028441	0.00006011	-4.73	<.0001
SVCAR	0.01891	0.00459	4.12	<.0001
GFCAR2	-0.00000506	0.00000766	-0.66	0.5090
GFCAR	-0.00515	0.00344	-1.50	0.1350
IBBCAR2	-0.00040164	0.00014068	-2.86	0.0045
IBBCAR	0.00938	0.00958	0.98	0.3281
DOMINANCECAR	-0.10565	0.02622	-4.03	<.0001
ERACAR	-0.09543	0.04229	-2.26	0.0245
WPCAR	0.01984	0.00661	3.00	0.0028
SVCAR2	-0.00002646	0.00001805	-1.47	0.1434
GCAR	0.00361	0.00119	3.03	0.0025
GCAR2	-0.00000329	9.729385E-7	-3.38	0.0008
POSSTART	0.22913	0.08515	2.69	0.0074
HRCAR	-0.00443	0.00184	-2.41	0.0162
LCAR2	0.00010333	0.00003974	2.60	0.0096
LCAR	-0.01316	0.00677	-1.94	0.0525

APPENDIX B. VARIABLES CONSIDERED

Table B1. Batting Statistics Considered

Batting Statistics **(Several squared statistics used with few significant in results)	
Games(G)	Career Games(GCAR)
Plate Appearances(PA)	Career Plate Appearances(PACAR)
At-Bats(AB)	Career At-Bats(ABCAR)
Runs(R)	Career Runs(RCAR)
Hits(H)	Career Hits(HCAR)
Doubles(DB)	Career Doubles(DBCAR)
Triples(TR)	Career Triples(TRCAR)
Home Runs(HR)	Career Home Runs(HRCAR)
Runs-Batted-In(RBI)	Career Runs-Batted-In(RBICAR)
Stolen Bases(SB)	Career Stolen Bases(SBCAR)
Caught Stealing(CS)	Career Caught Stealing(CSCAR)
Walks(BB)	Career Walks(BBCAR)
Strikeouts(SO)	Career Strikeouts(SOCAR)
Total Bases(TB)	Career Total Bases(TBCAR)
Ground-into-Double Play(GDP)	Career Ground-into-Double Play(GDPCAR)
Hit by Pitch(HBP)	Career Hit by Pitch(HBPCAR)
Sacrifice Hits(SH)	Career Sacrifice Hits(SHCAR)
Sacrifice Flies(SF)	Career Sacrifice Flies(SFCAR)
Intentional Walks(IBB)	Career Intentional Walks(IBBCAR)
Batting Average (BA)	Career Batting Average (BACAR)
On-Base-Percentage (OBP)	Career On-Base-Percentage (OBPCAR)
Slugging Percentage (SLG)	Career Slugging Percentage (SLGCAR)
C(1=Catcher, 0=Other)	SS(1=Shortstop, 0=Other)
1B(1=First Baseman, 0=Other)	OF(1=Outfielder, 0=Other)
2B(1=Second Baseman, 0=Other)	Bats_Left(1=Hits Left, 0=Other)
3B(1=Third Baseman, 0=Other)	Bats_Right(1=Hits Right, 0=Other)
Year(Year-2010)	

Table B2. Pitching Statistics Considered

Pitching Statistics **(Several squared statistics used with few significant in results)	
W(Wins)	WCAR(Career Wins)
L(Losses)	LCAR(Career Losses)
G(Games Pitched)	GCAR(Career Games Pitched)
GS(Games Started)	GSCAR(Career Games Started)
GF(Games Finished)	GFCAR(Career Games Finished)
CG(Complete Games)	CGCAR(Career Complete Games)
SHO(Shutouts)	SHOCAR(Career Shutouts)
SV(Saves)	SVCAR(Career Saves)
IP(Innings Pitched)	IPCAR(Career Innings Pitched)
H(Hits Allowed)	HCAR(Career Hits Allowed)
R(Runs Allowed)	RCAR(Career Runs Allowed)
ER(Earned Runs Allowed)	ERCAR(Career Earned Runs Allowed)
HR(Home Runs Allowed)	HRCAR(Career Home Runs Allowed)
BB(Walks Allowed)	BBCAR(Career Walks Allowed)
IBB(Intentional Walks Allowed)	IBBCAR(Career Intentional Walks Allowed)
SO(Strikeouts)	SOCAR(Career Strikeouts)
HBP(Hit Batter)	HBPCAR(Career Hit Batter)
BK(Balks Allowed)	BKCAR(Career Balks Allowed)
WP(Wild Pitches)	WPCAR(Career Wild Pitches)
BF(Batters Faced)	BFCAR(Career Batters Faced)
ERA(Earned Run Average)	ERACAR(Career Earned Run Average)
DOMINANCE (SO/9*IP)	DOMINANCECAR (SOCAR/9*IPCAR)
CONTROL (BB/9*IP)	CONTROLCAR(BBCAR/9*IPCAR)
COMMAND (SO/BB)	COMMANDCAR (SOCAR/BBCAR)
WHIP ((BB+H)/IP)	WHIPCAR ((BBCAR+HCAR)/IPCAR)
Year(Year-2010)	POS(1=Staring Pitcher, 0=Relief Pitcher)

APPENDIX C. SAS CODE

```
*Yearly Models; *Position Players;
proc import file='F:/Thesis/Thesis New/Samples/Batting_Position_1.xlsx' out=Yearly_Batting
dbms=xlsx replace;
run;
*Model;
proc reg data=yearly_batting;
        model lnsalary = al year1 POS1B POS2B POS3B POSC POSSS POSDH batsright
        batsleft G PA AB R H  DB TR HR RBI SB CS BB
        SO TB GDP HBP SH SF IBB BA OBP SLG G2 DB2 HR2 SB2 BB2 GDP2 HBP2
        IBB2/selection=stepwise;
run;
*Which corresponds to;
proc reg data=yearly_batting outest=Yearly_Bat_Estimates press;
        model lnsalary = RBI TR G2 G PA SH SO POSC POS1B POS2B POS3B POSSS
        POSDH DB DB2;
        output out=yearly_predictions P=Pred_Values R=Resid_Values;
run;
*Pitchers Yearly Model;
data Pitching;
        length name $30.;
        infile 'F:/Thesis/Thesis New/Final Modified Position Pitchers.csv' dlm=',' firstobs=2 dsd;
        input year      Lg $    name    $ Team $        W       L       G       GS      GF      CG
        SHO     SV      IP      H       R       ER      HR      BB      IBB
        SO      HBP     BK      WP      BF      ERA     COMMAND         DOMINANCE
        CONTROL         HR9     WHIP    WRIP    ERACAR          COMMANDCAR
        DOMINANCECAR
        CONTROLCAR              HR9CAR          WHIPCAR         WRIPCAR         BBCAR
        BFCAR           BKCAR           CGCAR           ERCAR           GCAR    GFCAR
        GSCAR           HCAR
        HBPCAR          HRCAR           IBBCAR          IPCARLCAR       RCAR    SHOCAR
        SOCAR           SVCAR           WCAR            WPCAR           salary  throws $ POS  $;
run;
*Model;
proc reg data=Pitching;
        model lnsalary = W L year1 G GS GF CG SHO SV IP H R ER HR BB IBB SO HBP BK
        WP BF ERA Command Dominance Control HR9 WHIP WRIP AL POSSTART
        throwsright W2 L2 G2 GS2 GF2 CG2 SHO2 SV2 IP2 H2 R2 ER2 HR2 BB2 IBB2 SO2
        HBP2 BK2 WP2 BF2 /selection=stepwise;
run;
*THIS MODEL WOULD CORRESPOND TO:;
proc reg data=Pitching outest=Yearly_Pitch_Estimates;
        model lnsalary = SV Year1 GS BB2 SV2 BK2 CG2 L BK CG;
        output out=pitchpred_year cookd=cooks P=Pred_year R=Resid_year;
run;
```

```
*Exlude Extreme Observation;
proc sort data=pitchpred_year;
        by descending cooks;
run;
data Pitching_1;
        set Pitching;
        if name = 'JamesShields' and year = 2011 then delete;
run;
proc reg data=Pitching_1 outest=Yearly_Pitch_Estimates;
        model lnsalary = SV Year1 GS BB2 SV2 BK2 CG2 L BK CG;
        output out=pitchpred_year cookd=cooks P=Pred_year R=Resid_year;
run;
*Career Models;
*Position Players;
proc import     file='F:/Thesis/Thesis/New/Samples/Batting_Career_Position_1.xlsx'
out=Career_Batting dbms=xlsx replace;
run;
*Model;
proc reg data=career_batting;
        model lnsalary = AL year1 POS1B POS2B POSSS POS3B POSC POSDH batsright
        batsleft BACAR OBPCAR SLGCAR ABCAR BBCAR CSCAR DBCAR GCAR
        GDPCAR HCAR HBPCAR HRCAR IBBCAR PACAR RCAR RBICAR SBCAR
        SFCAR SOCAR TBCAR TRCAR BBCAR2 CSCAR2 DBCAR2 GCAR2
        GDPCAR2 HCAR2 HBPCAR2 HRCAR2 IBBCAR2 RCAR2 RBICAR2 SBCAR2
        SFCAR2 SHCAR2 SOCAR2 TBCAR2 /selection=stepwise;
run;
*Which corresponds to:
proc reg data=career_batting outest=Career_Bat_Estimates press;
        model lnsalary = TBCAR GCAR2 GCAR TBCAR2 SHCAR2 SHCAR POS1B POS2B
        POSSS POS3B POSDH POSC CSCAR2 CSCAR RCAR2 RCAR GDPCAR2 GDPCAR
        ABCAR SBCAR year1;
        output out=car_pred cookd=cooks P=Pred_Values R=Resid_Values;
run;
proc sort data=car_pred;
        by descending cooks;
run;
*Exclude extreme observations;
data career_batting_2;
        set career_batting;
        if name = 'GaryMatthews' and year = 2010 then delete;
        if name = 'IvanRodriguez' and year = 2011 then delete;
        if name = 'KenGriffey' and year = 2010 then delete;
        if name = 'IvanRodriguez' and year = 2010 then delete;
run;
```

```
*Final model with excluded players;
proc reg data=career_batting_2 outest=Career_Bat_Estimates press;
        model lnsalary = TBCAR GCAR2 GCAR TBCAR2 SHCAR2 SHCAR POS1B POS2B
        POSSS POS3B POSDH POSC CSCAR2 CSCAR RCAR2 RCAR GDPCAR2 GDPCAR
        ABCAR SBCAR year1;
        output out=car_pred cookd=cooks P=Pred_Values R=Resid_Values;
run;
*Pitchers;
data Pitching;
        length name $30.;
        infile 'F:/Thesis/Thesis New/Final Modified Position Pitchers.csv' dlm=',' firstobs=2 dsd;
        input year      Lg $    name    $ Team $        W       L       G       GS      GF      CG
        SHO     SV      IP      H       R       ER      HR      BB      IBB
        SO      HBP     BK      WP      BF      ERA     COMMAND DOMINANCE
        CONTROL         HR9     WHIP    WRIP    ERACAR          COMMANDCAR
        DOMINANCECAR
        CONTROLCAR              HR9CAR          WHIPCAR         WRIPCAR         BBCAR
        BFCAR           BKCAR           CGCAR           ERCAR           GCAR GFCAR
        GSCAR           HCAR
        HBPCAR          HRCAR           IBBCAR          IPCARLCAR       RCAR    SHOCAR
        SOCAR           SVCAR           WCAR            WPCAR           salary  throws $ POS  $;
run;
*Model;
proc reg data=Pitching;
        model lnsalary = AL POSSTART throwsright year1 WCAR LCAR GCAR GSCAR
        GFCAR CGCAR SHOCAR SVCAR IPCAR
        HCAR RCAR ERCAR HRCAR BBCAR IBBCAR SOCAR HBPCAR BKCAR WPCAR
        BFCAR ERACAR CommandCAR DominanceCAR ControlCAR HR9CAR WHIPCAR
        WRIPCAR LCAR2 GCAR2 GSCAR2 GFCAR2 CGCAR2 SHOCAR2 SVCAR2
        IPCAR2 HCAR2 RCAR2 HRCAR2 BBCAR2 IBBCAR2 SOCAR2 BKCAR2
        WPCAR2/selection=stepwise;
run;
*THIS MODEL CORRESPONDS TO:;
proc reg data=Pitching outest=Career_Pitch_Estimates press;
        model lnsalary = SOCAR SOCAR2 BBCAR YEAR1 WPCAR2 SVCAR GFCAR2
        GFCAR IBBCAR2 IBBCAR DOMINANCECAR ERACAR WPCAR
        SVCAR2 GCAR             GCAR2 POSSTART HRCAR LCAR2 LCAR;
        output out=car_pitch cookd=cooks P=Pred_Values R=Resid_Values;
run; *R2 equals 0.7096;
proc sort data=car_pitch;
        by descending cooks;
run;
```

```
*Exclude extreme observations;
data Career_Pitching_3;
        set Pitching;
        if name = 'MarianoRivera' and year=2011 then delete;
        if name = 'BillyWagner' and year = 2010 then delete;
run;
proc reg data=Career_Pitching_3 outest=Career_Pitch_Estimates press;
        model lnsalary = SOCAR SOCAR2 BBCAR YEAR1 WPCAR2 SVCAR GFCAR2
        GFCAR IBBCAR2 IBBCAR DOMINANCECAR ERACAR WPCAR
        SVCAR2 GCAR      GCAR2 POSSTART HRCAR LCAR2 LCAR;
        output out=car_pitch cookd=cooks P=Pred_Values R=Resid_Values;
run;
```