University of Nevada, Reno

Analysis of Statistical Impact of Steroids in Professional Major League Baseball Players

A thesis submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Science in Mathematics and the Honors Program

by

Joseph L. Ward III

Professors Thomas Quint & Ilya Zaliapin, Thesis Advisors

May, 2015

UNIVERSITY OF NEVADA RENO

THE HONORS PROGRAM

We recommend that the thesis
prepared under our supervision by

JOSEPH L. WARD III

entitled

Analysis of Statistical Impact of Steroids in Professional Major League Baseball Players

be accepted in partial fulfillment of the
requirements for the degree of

BACHELOR OF SCIENCE, MATHEMATICS – STATISTICS OPTION

_____
Thomas Quint, Ph.D., Thesis Advisor

_____
Ilya Zaliapin, Ph.D., Thesis Advisor

_____
Tamara Valentine, Ph. D., Director, **Honors Program**

May, 2015

Abstract

Rampant steroid usage tainted Major League Baseball (MLB) in the late 1980s, and decades later, steroid usage is still a serious issue. Steroids, along with other illegal substances, have heavily impacted various statistics in professional baseball (Petersen, Jung, & Eugene Stanley, 2008). Many records—a more notable one being Barry Bond's 73 homerun season—have been broken during this timeframe, which has been coined as the "Steroid Era" of baseball (Rymer, 2012). In a sport with more statistics than any other, the impact steroid usage has on baseball statistics becomes fascinating, and this impact can be mapped in a variety of ways. In fact, there is an entire branch of statistical modeling specifically for baseball formally known as "sabermetrics" (SABR).

The core of this thesis is an attempt to analyze the statistical impact steroids have on baseball statistics at a professional level. By utilizing various baseball sabermetrics to collect data, this study examines how steriods have a career-wide impact on the statistical distribution of a MLB player who used steroids with respect to a player who refrained from usage of such illegal substances. By applying various analyses on said data, potential differences can be made quantifiable. These results could be telling enough to portray suggestive anomalies in a MLB player's statistics. On a larger level, these results could be telling enough to discourage steroid usage among professional baseball players entirely.

Acknowledgements

Firstly, I would like to thank my my mentors Dr. Thomas Quint and Dr. Ilya Zaliapin, who were both willing to give my thesis topic a chance. Quint provided me with a much appreciated baseball-oriented point of view that helped me produce my methodology, and Zaliapin's strength in statistical analysis greatly helped me when I compiled my results. I would also like to thank the Honors program—in particular, Dr. Tamara Valentine—for providing me with the opportunity to produce this thesis. Additionally, I would like to thank the fellows of hallofstats for producing the value-based similarity score generator that I utilized for my thesis. I truly belive that this resource helped me produce cleaner results, and for that I am grateful. Lastly, I thank my father, Joseph Leo Ward Jr., for instilling in me the love of baseball, for looking at my work, and for letting me bounce all sorts of ideas for my thesis off of him.

Table of Contents

List of Tables

List of Figures

Introduction

Out of all of the sports to garner popularity in America, only baseball has been labeled as America's pastime, and for good reason. The sport existed at a professional level with professional teams as early as 1869, and it has constantly developed since then (Rader, 1992). Predictably, baseball has evolved over the years and, as with any sport, so have the performance levels of its athletes. While the performance progression by typical Major League Baseball (MLB) athletes can be expected over time, recently there has been a highly significant rise in certain baseball players' statistics over the past two decades, such as home runs (Petersen, Jung, & Eugene Stanley, 2008).[1] Despite an expectation of some increase in players' performance levels over the years, this modern-day surge is unnatural. Given the history of baseball, the recent upswing in the players' performance season-wide statistics is likely attributable to the players' use of performance enhancing drugs or steroids. This contemporary period is often labeled as the "Steroid Era" of baseball.

The "Steroid Era" of baseball began during the 1980s, arguably due to simple economic incentive. In 1980, the minimum salary of a MLB player surpassed the mean US household income, and by 1990, the minimum salary of a MLB player surpassed $100,000.00 (Boss, 2012). In fact, studies suggest that steroid usage has the potential to increase yearly salary of the modern MLB player by about 2 million dollars per season. With the average career length of a MLB player being 6 years, it makes sense for athletes to take a gamble on steroids (Lenhardt, 2010). Many players have taken this gamble, and this shows through the exceedingly high single season homerun records set by many baseball players, including Mark McGwire (70 homeruns), Barry Bonds (73 homeruns), and Sammy

[1] Unless otherwise noted, "players" or "athletes" shall mean baseball players.

Sosa (66 homeruns). The Mitchell Report – an investigative report by George Mitchell to the Commissioner of Baseball on the usage of illegal substances – brought awareness to the MLB community of the propensity of steroid usage among baseball players (Mitchell, 2007).

The Mitchell Report cast a spot light on the issue of whether players' use of controlled substances, including amphetamines, human growth hormones (HGH), anabolic steroids and testosterone, explain the recent unusual inflated performance statistics during this "Steroids Era". The report showed that steroid usage was not uncommon in MLB, and that various players have purchased and or used steroids throughout their careers (Mitchell 2007). Although stricter testing policies and punishments have been established, steroid usage has remained a prevalent issue in America's favorite pastime (Vinton, 2014).

This issue begs the question whether or not there are ways to reveal players' steroid usage besides random drug testing. In light of the evolution of performance statistics for MLB players, by analyzing career-wide statistics of MLB players connected with steroid usage and MLB players who are not and comparing these groups, it is possible to discover statistical trends that suggest which players are using illegal substances. Specifically, the aforementioned analysis focuses on Wins Above Replacement (WAR) and value-based similarity scores—methodologies branching from Bill James' sabermetrics.[2] Using these similarity scores, I grouped baseball players into three different groups: players who have been caught using illegal substances, players who have not, and players who played prior

[2] Sabermetrics is the term for the empirical analysis of baseball. It is derived from the acronym SABR, which stands for the Society for American Baseball Research, as coined by baseball analyst Bill James. *See* Lewis, Michael M. (2003). *Moneyball: The Art of Winning an Unfair Game*. New York: W. W. Norton. ISBN 0-393-05765-8.WAR represents the number of wins this player contributed, above what a replacement level hitter, fielder, and pitcher would have done. This is addressed in greater depth below, at "Literature Review."

to the aforementioned "Steroid Era." I then performed regression and threshold analysis on these groups to find any evident differences between the groups.

This research has potential for significance in that it helps give a new, quantifiable perspective on how steroids influence athletes' performances and their careers in terms of statistics. Rather than focus specifically on homerun production, I took various sabermetrics into account and focused primarily on the rates of changes of MLB players' statistics throughout their careers to see if steroids affected career longevity or influenced the rate at which players' skills declined as they aged. Furthermore, while there is other research that focuses on WAR as well as aging, I have not come across any research that employs a similarity score-based methodology like the one incorporated in this thesis. For example, I came across a similar study by Furnald (2012) that analyzed age with respect to WAR, but it utilized dummy variables to generate various test groups rather than similarity scores. Furnald found that steroid-using players peaked later in their careers, yet their skills declined more rapidly after this peak. I believe that although my methodology differs significantly, that it provided clean feasible data, and that this data illustrates differences between MLB players who play baseball cleanly and MLB players who resort to using illegal substances. My hypothesis was that athletes who have not used steroids, or those who have not been sanctioned, will have a more even distribution of WAR in comparison to sanctioned athletes who have used controlled performance enhancing substances, who will have a less balanced distribution. In addition, I expected to find that steroid using athletes would peak later on in their careers and have longer lasting careers. My results contrasted my hypothesis in that the regressions of the three groups appeared to be quite similar, although there were differences in variation that support my hypothesis. The steroid

group had highest variation, and this was detectable in various Figures. Furthermore, threshold analysis showed that during/post-steroid era players were most consistent for lower values of WAR, yet had smaller maximum WAR values. Due to the nature of these results, this thesis could potentially encourage players to abstain from using illegal substances in ways that random testing cannot. By mapping a current MLB player's statistics against my results, telling anomalies could be discovered that might otherwise be ignored.

Literature Review

In the 1980s, prior to the Steroid Era, Bill James coined the term "sabermetrics", which is defined as "the search for objective knowledge about baseball" (SABR). Simply put, sabermetrics is a synonym for baseball statistics. Although this terminology is fairly young, sabermetrics has been around for decades, and it is much more sophisticated than the early tabulation of basic box score statistics and hits from which sabermetrics evolved from (SABR). For my thesis, I am capitalizing on this evolution of baseball statistics by focusing on similarity scores. Similarity scores are a more advanced baseball statistic that measure similarity between two players by comparing their career-wide statistics of games played, at bats, runs scored, hits, doubles, triples, homeruns, and more (Baseball Reference). However, usage of similarity scores in research is quite limited, as there is a fair amount of controversy regarding the effectiveness of similarity scores. Specifically, some baseball statisticians critique the accuracy similarity scores provide. First, similarity scores can be less than reliable when honing in on players with shorter than average careers. In other words, the longer a player's career, the more meaningful his similarity

scores are. Secondly, similarity scores are not entirely reflective of the era in which a MLB athlete played. For example, a player hitting 20 or more homeruns is much more impressive during the dead ball era versus the steroid era, and traditional similarity scores fail to encapsulate this concept (Waters, 2008). However, the concept of similarity scores is a novel one, to the point that various alternative similarity scores have been created. One of these alternative similarity scores, created by hallofstats, is known as a "value-based similarity score". Rather than focusing on raw data (like traditional similarity scores), hallofstats's value-based similarity scores utilize WAR or Wins Above Replacement and WAA or Wins Above Average.

WAR, the acronym for "Wins Above Replacement," can be thought of as an all inclusive statistic with this example: "If [a] player got injured and their team had to replace them with a freely available minor leaguer or a AAAA player from their bench, how much value would the team be losing?" (What is WAR?). To further expand, if a playerhad a seasonal WAR of 3.3, then he is worth 3.3 wins to his team during that season. While WAR may have its imperfections—there are multiple variations of calculating WAR—it focuses on players' values relative to the season they are playing in and can be a more accurate indicator of MLB players' true skill. Furthermore, WAR is a much more telling statistic. For example, if we focused on two MLB players, and it was known that MLB player #1 hit 20 homeruns last season and that MLB player #2 had a WAR of 5.0, more could be derived about player #2. As promising as 20 homeruns sounds, it could very well be that player #1 underperforms in different ways. For example, player #1 might make a lot of defensive errors, ground into double plays (GIDP) frequently, or have a low on base percentage. On the other hand, player #2 is guaranteed to have an overall positive value to his team.

WAA (Wins Above Average) is essentially the same stat as WAR, but in contrast, WAA is calculated without a replacement adjustment based on playing time. In other words, a player is not rewarded for "being there" (Darowski, 2012). Simply put, if two MLB players contributed equally to their respective teams, but one player played significantly less than the other, WAA (unlike WAR) would capture this difference of playing time. In conclusion, because these value-based similarity scores are calculated with sabermetrics instead of "raw data" (batting average, hits, runs, homeruns, etc.), generating similarity scores using value-based similarity scores would serve as a more efficient methodology for my thesis.

Just as statistical analysis has evolved in baseball, so has the usage of performance enhancing drugs. Going back to ancient Olympic/Roman times, athletes resorted to taking herbs and mushrooms in an attempt to maximize their performance in their sports. Performance enhancing drugs continued to evolve, and in 1928, they were first banned from track and field events (Grossman, Kimsey, Moreen, & Owings, 2012). Eventually, steroids—from human-growth hormone and testosterone to amphetamines—were banned from MLB in 1991, but players were not tested for steroids until 2003 (Carise, 2013). Once testing began, offensive statistics began to drop across the board. The once high, league leading, single season homerun values of 66, 70, & 73 homeruns accomplished by players Sosa, McGwire and Bonds dropped drastically to the 40-50 range over the following years (Rymer, 2013). Despite this decrease in offensive power production, the MLB drug policy continued to become stricter and suspensions enforced at a steady rate from 2005 to present (The Steroids Era).

After review of the literature of baseball/steroid related studies, it is evident that some studies are similar to this study. Furnald (2012) performed a similar analysis of graphing age versus WAR of MLB players. His other samples, however, are not derived from similarity scores. Instead, a "dummy variable" is utilized. His results show that MLB players who use steroids perform slightly better in terms of WAR and peak at a later age, but decline athletically at a faster rate. He partially attributed this faster decline to the fact that steroids led to excess muscle gain to the point that joints and ligaments could not handle these change, thereby the use of steroids lead to more injuries. In contrast, various MLB players who resorted to steroids used them solely to recover from injuries more quickly. Slugger Mark McGwire and ace pitcher Andy Pettitte are solid examples of player who resorted to steroids to recover from injury (McGwire comes clean, 2010). Similarly, former MLB player and steroid user Troy Glaus claimed to use steroids in an attempt to recover more quickly from an injury, and he claimed that overall, his use of steroids was a "blessing in disguise" since he became healthier and more educated about his health and nutrition (Singer, 2009). Even though the relationship between a MLB player and steroid usage varies, it is clear that steroids have influenced baseball statistics.

Methodology

My first step was to gather compile a feasible list of players who have either been directly or indirectly involved in steroid scandals. I did some research and compiled a massive list of all MLB players who qualify for this list, whether indirectly (Mitchell Reports) or directly (tested positive for illegal substances and served a suspension). With

all of the indirectly involved candidates, I judged players individually, and I focused on two criteria.

First, I focused on whether or not the player had an alibi or conversely, if they confessed to the transgression. For example, if a player was mentioned in the Mitchell Report for having above average testosterone levels, but he had testicular cancer and was required to receive testosterone injections; and his conditions and treatments were documented by his doctor, then he would be exonerated from my list of steroid using players. However, many players admitted post-career that they did, indeed, use some sort of illegal substance, including but not limited to amphetamines, anabolic steroids, HGH, and testosterone.

Second, I looked at a player's overall career and verified that he played at least five full seasons or the equivalent of five full seasons. Different events such as injuries, being called up from the minors, leaving for war, or coming out of retirement could explain gaps in a player's career. Since I am measuring change over time, and a longer timeframe is preferred, if a player did not play for an amount equivalent to five full seasons, then I concluded that his data (seasonal WAR values) would be insufficient. Overall, my list totals 78 players, 50 of whom are hitters. Due to the majority of these players being hitters, this thesis focuses primarily on hitters.

My next step was to calculate the top value-based similarity scores for each player on this list of players who used performance-enhancing substances. Specifically, I found the most statistically similar player for each hitter of this "performance-enhancing substance" list who played prior to the steroid era. I repeated this process for each hitter who played during/after the steroid era—and did not use illegal substances. Rather than

use Bill James' traditional similarity scores, I used an alternative method for calculating these scores. Specifically, I used hallofstats's value-based similarity scores generator that takes the following criteria into account:

- WAR Batting Runs
- WAR Baserunning Runs
- WAR Double Play Runs
- WAR Defensive Runs
- WAR Positional Runs
- WAR Pitching Runs
- adjWAR
- adjWAA
- Plate Appearances
- Innings Pitched

(Darowski, Chupp, & Berkowitz). The calculated similarity scores ranged from 0 – 250, with the exception of similarity scores generated from atypical players such as Barry Bonds, whose top similarity scores exceeded this threshold. An example of 3rd baseman Josh Donaldson's similarity scores, generated from hallofstats is cited in Table 1. Note that this is an arbitrary example. The numerical values denote how statistically similar each player's career is with respect to Josh Donaldson. Larger similarity scores represent larger statistical difference between Josh Donaldson and the respective player, and smaller scores indicate lower statistical difference. Therefore, it can be seen in Table 1 that Morgan Ensberg is the most statistically comparable player to Josh Donaldson at a career-wide, value-based level.

Morgan Ensberg	98
Salvador Perez	105
Todd Frazier	105
Tim Naehring	106
Matt Carpenter	107
Jonathan Lucroy	111
Anthony Rizzo	114
Alex McKinnon	115
Aaron Robinson	115
Adolfo Phillips	118

Table 1 – Sample Value-Based Similarity Score Chart: This chart is a representation of the top 10 players most similar to 3rd baseman Josh Donaldson, as generated per the hallofstats value-based similarity score generator.

Once I derived all of the players' respective top similarity scores, I plan separated the players into two groups. One included players that played prior to the steroid era, and the other consisted of players that played the majority of their careers during/after the Steroid Era. For purposes of maintaining accuracy for this portion of the methodology, I used 1980 as a cutoff year when generating these two groups. The third group is the initial group of players who used performance-enhancing substances.

After these three groups were established, I used Minitab to collect age-based seasonal data (WAR) from each player, graphed the average WAR versus age for each of these three groups, performed regression analysis for each group, and tested variance between the three groups. This regression analysis was done with respect to the regression model:

$$\mathbf{Z = a*A^2 + b*A + c + \epsilon}$$

In this model, **a**, **b,** and **c** are real constants (respectively the regression coefficients and the y-intercept), **A** is the independent variable, **Z** is the dependent variable, and **ϵ** is the error term (Meerschaert 2013). For the purpose of my thesis, **Z** represents WAR, **A** represent s age, and my regression-based graphs show age (x-axis) versus WAR (y-axis). After using Minitab to graph this relationship for the three groups, I originally intended to use ANOVA analysis (analysis of variance) to analyze differences between the three groups and thereby measure the impact of steroids on MLB players in terms of WAR. However, based on the results, I deemed other methodologies more efficient. Specifically, I created a computer program using C++ that performed a threshold analysis. This program[3] reads every value and measures the percentage of values that exceed thresholds from -3.5 to 14. In addition, I measured different thresholds, variance, variation, and various ranges using boxplots.

Results

Regression analysis of the data that were collected for the three groups did not show much difference between the three groups, which was contrary to my hypothesis. Below is the graph of the data for all three groups combined (see Figure 1), as well as the graphs for each individual group for easier viewing (see Figures 2, 3, and 4). Figure 1 is a visual representation of change in WAR with respect to age. The x-axis represents age, and the y-axis represents WAR. For example, the blue point (36, 11.9) means that one of the players in the steroid group (in this case Barry Bonds) had a WAR value of 11.9 during the season in which he was 36 years old. The lines are regression fits, which are lines that best fit all of the corresponding data. We see here that these regression fits follow a negative

[3] The code utilized for this program can be found on page 30.

quadratic pattern, and this is due to the fact that athletes generally continue to statistically improve throughout their careers until they physically peak.

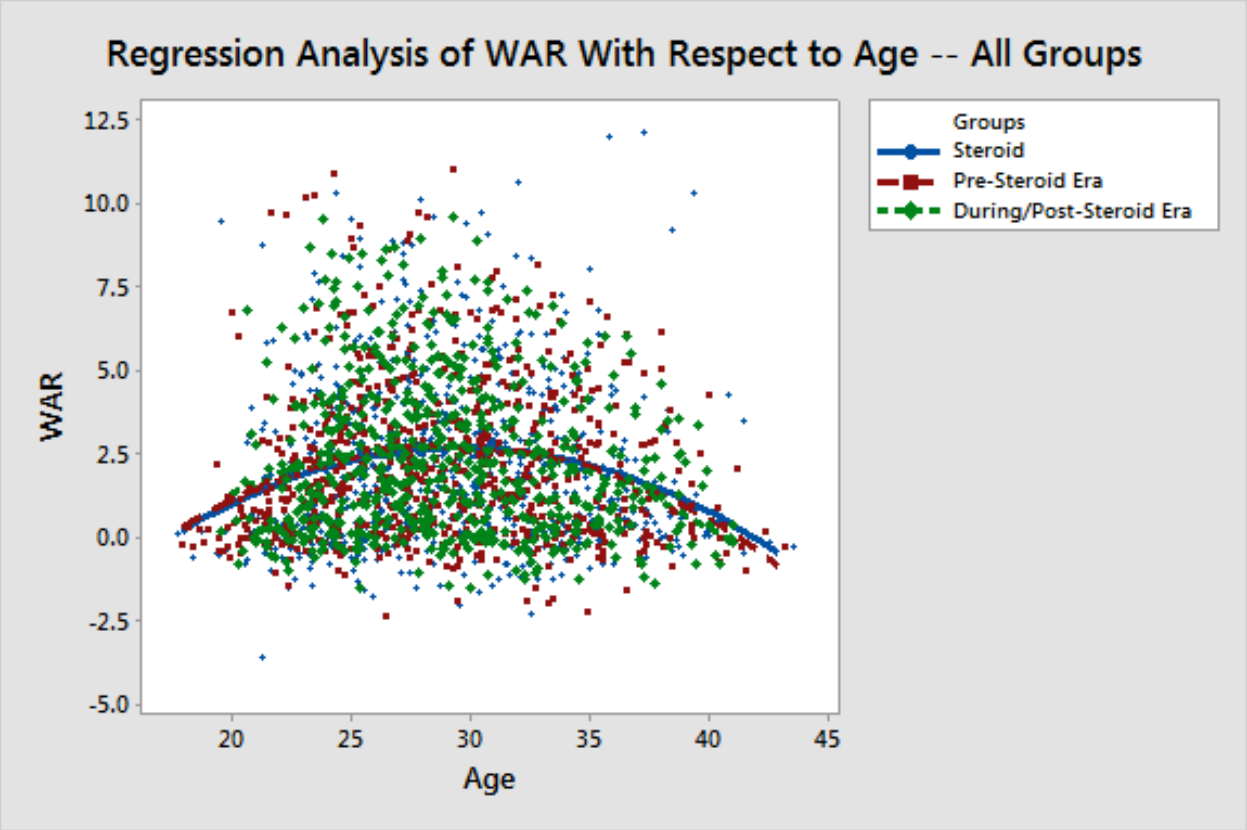

Figure 1 – Graph of Regression Analysis of WAR With Respect to Age – All Groups: This graph is a visual representation of change in WAR with respect to age. The x-axis represents age, and the y-axis represents WAR. Due to the closeness of the regression fits on this graph, these three groups are separated—with equal scaling—for easier viewing in Figures 3, 4, and 5.

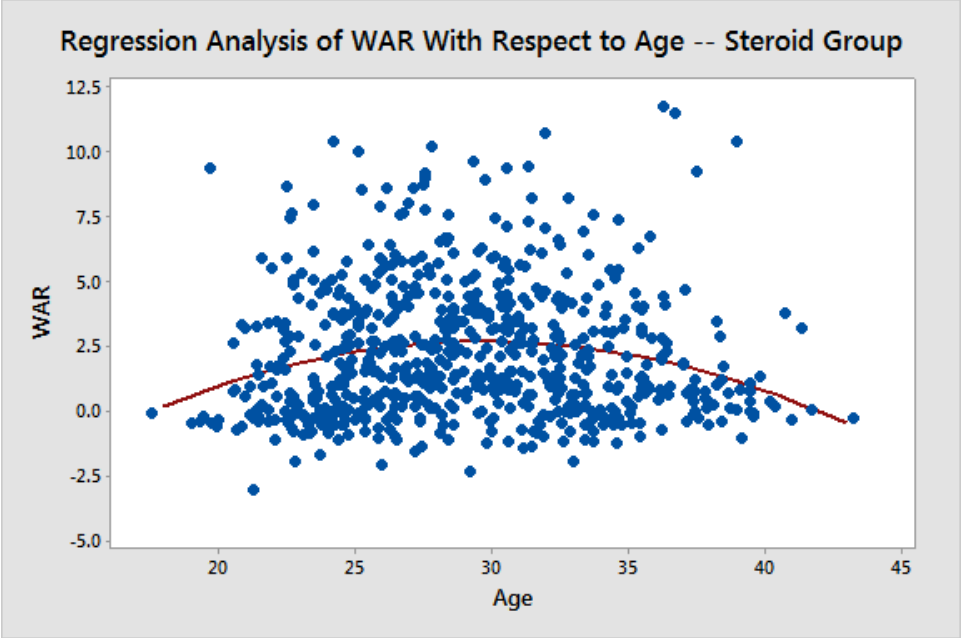

Figure 2 – Graph of Regression Analysis of WAR With Respect to Age – Steroid Group: As mentioned in figure 1, this figure is a visual representation of change in WAR with respect to age. However, this figure only represents the steroid group of MLB players. Once again, the x-axis represents age, and the y-axis represents WAR.

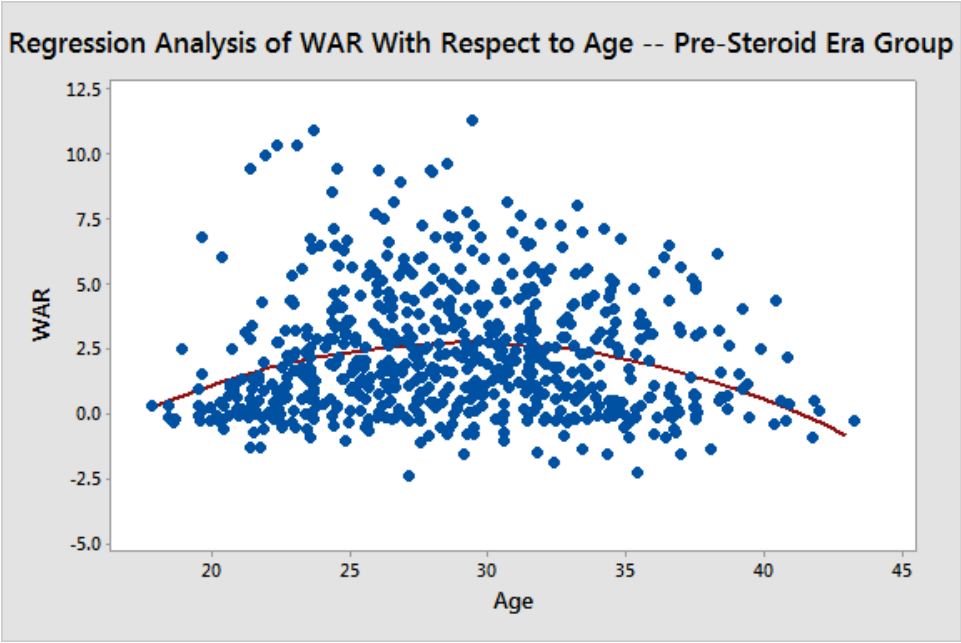

Figure 3 – Graph of Regression Analysis of WAR With Respect to Age – Pre-Steroid Era Group: As mentioned in figure 1, this figure is a visual representation of change in WAR with respect to age. However, this figure only represents the pre-steroid era group of MLB players. Once again, the x-axis represents age, and the y-axis represents WAR.

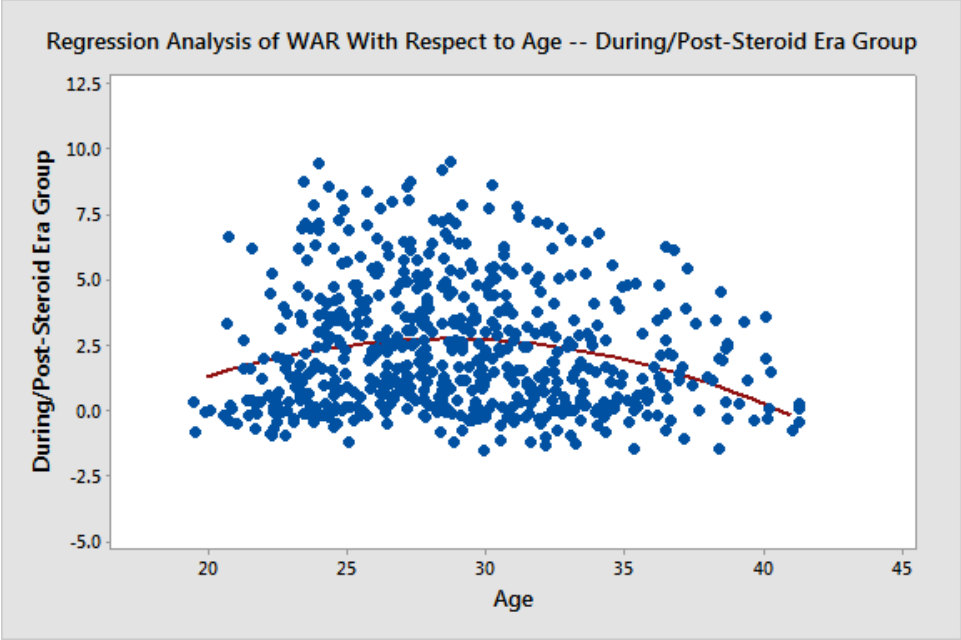

Figure 4 – Graph of Regression Analysis of WAR With Respect to Age – During/Post-Steroid Era Group: As mentioned in figure 1, this figure is a visual representation of change in WAR with respect to age. However, this figure only represents the during/post-steroid era group of MLB players. Once again, the x-axis represents age, and the y-axis represents WAR.

Contrary to my expectations, the graphs produced from the regression analysis of the three groups appear to be similar in nature. It can be seen that the maximums, or peaks of the three graphs are all very close both in terms of age and value (see Table 2).

Maximums of the Regression Fits of Each Group			
	Age	WAR	R-Sq.
Steroid Group	29.75	2.65	3.7%
Pre-Steroid Era Group	29.32	2.71	5.8%
During/Post Steroid Era Group	28.68	2.72	4.7%

Table 2 – Chart of Maximums of the Regression Fits of Each Group: This chart contains values derived from the previous graphs (figures 1-4).

In the Age column, we see that the age at which players peak, and in the WAR column we see the associated WAR value for that age. Upon observation, it can be seen that age declines very slightly between each group, whereas WAR remains nearly constant. The

R-Sq. column represents the R^2 value. This is a value ranging from 0 to 100 percent (or 0 to 1) which represents how well the regressions fit the data. The higher the percentage, the better the graph matches with the data. Therefore, these R^2 values, which were obtained from the graphs in figures 1-4, are quite telling. The steroid group has the lowest R^2 score at 3.7 percent, the pre-steroid era group has the highest R^2 score at 5.8 percent, and the R^2 score of the during/post steroid era group is a near average of these at 4.7 percent. Due to the highly similar nature of the three graphs, these R^2 values suggest that there is the most variability in the steroid group, the least variability in the pre-steroid era group, and an intermediate amount of variability in the during/post steroid era group. To further confirm this finding, other statistics were used to measure these differences in WAR distribution. Figures 5 and 6—are boxplot figures that respectively illustrate the variance and variation of the WAR from the three groups (see Figures 5 & 6).

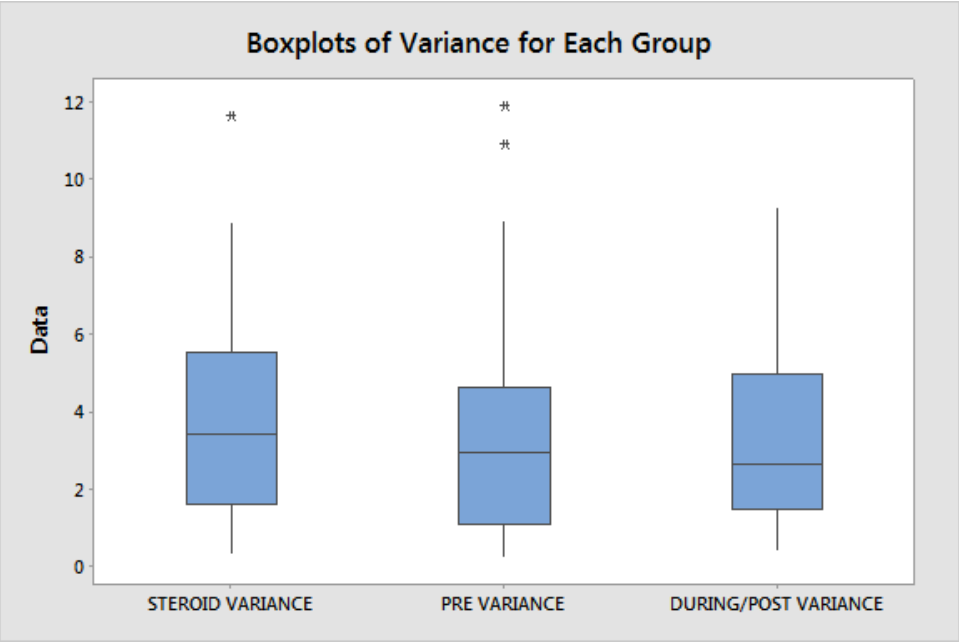

Figure 5 – Boxplots of Variance for Each Group: This diagram consists of three boxplots, one for the steroid group, one for the pre-steroid era group, and one for the during/post-steroid era group. Specifically, these boxplots represent the variance of WAR values of each MLB player in each group.

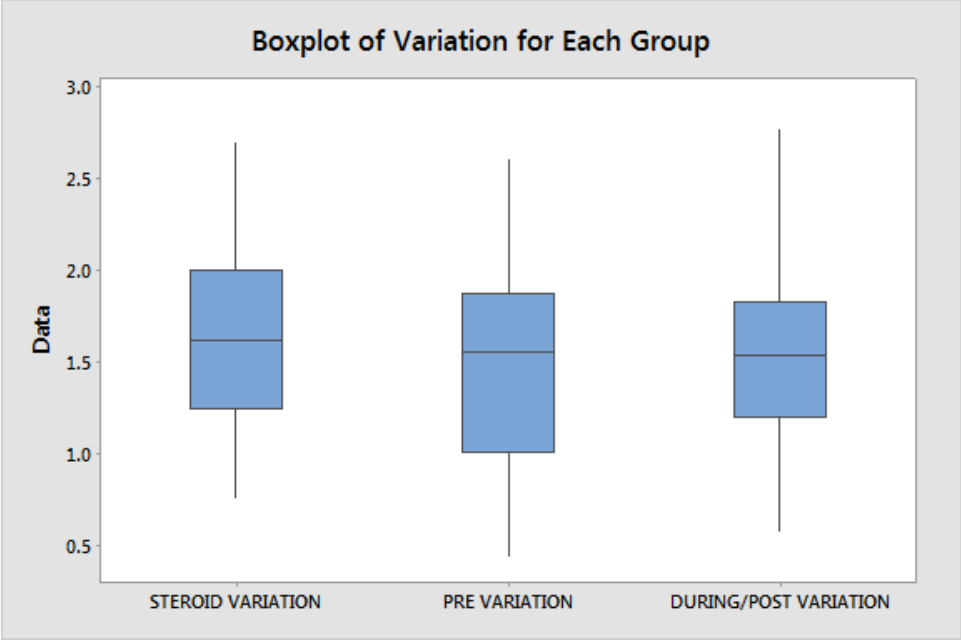

Figure 6 – Boxplots of Variation for Each Group: This diagram consists of three boxplots, one for the steroid group, one for the pre-steroid era group, and one for the during/post-steroid era group.

Despite the outliers, Figure 5 suggests higher levels of variance in the steroid group. In other words, the WAR values in the steroid group deviate the most from the mean out of the three groups. The statistic utilized in Figure 6 is somewhat similar, but rather than measure deviation, it measures the overall change of WAR values per season. For example, if a player played four seasons total and had WAR values of 1, 5, 5, and 1 in each season, his calculated variation would be equal to: (|(1-5)|+|(5-5)|+|(5-1)|)/4, which equals 2.0. We see once again that the steroid group has the most variation of the three groups. Although these values were supportive of the hypothesis, other analyses were utilized, one of them being threshold analysis.

As seen earlier in Figures 1-4, the thresholds for values for the different groups highly differed between the three groups. What was most noticeable was that unlike the other two groups, the during/post-steroid era group entirely lacked any values that exceeded the threshold of 10 WAR. Therefore, my first step in additional analysis was to

focus on threshold. Initially, I made a program using C++ (which can be found in the appendix) that analyzed the percentage of values that exceeded various thresholds, and I graphed the results (see Figure 7). On a similar vein, I also produced threshold-based boxplots. These plots represent percentages of each player's career in which each player's WAR exceeded the corresponding threshold (see Figures 8-10).

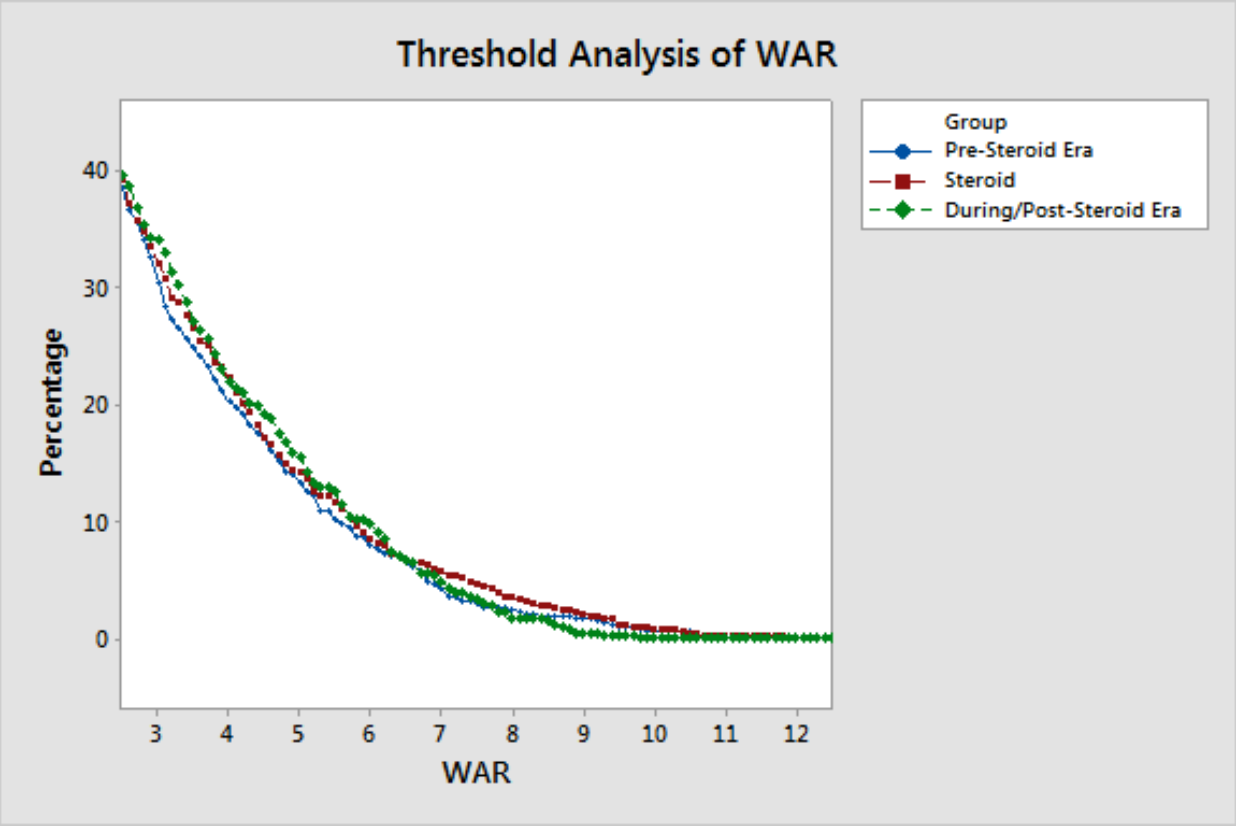

Figure 7 – Threshold Analysis of WAR: This chart is a visual representation of the percentage of values (x-axis) exceeding the WAR threshold (y-axis).

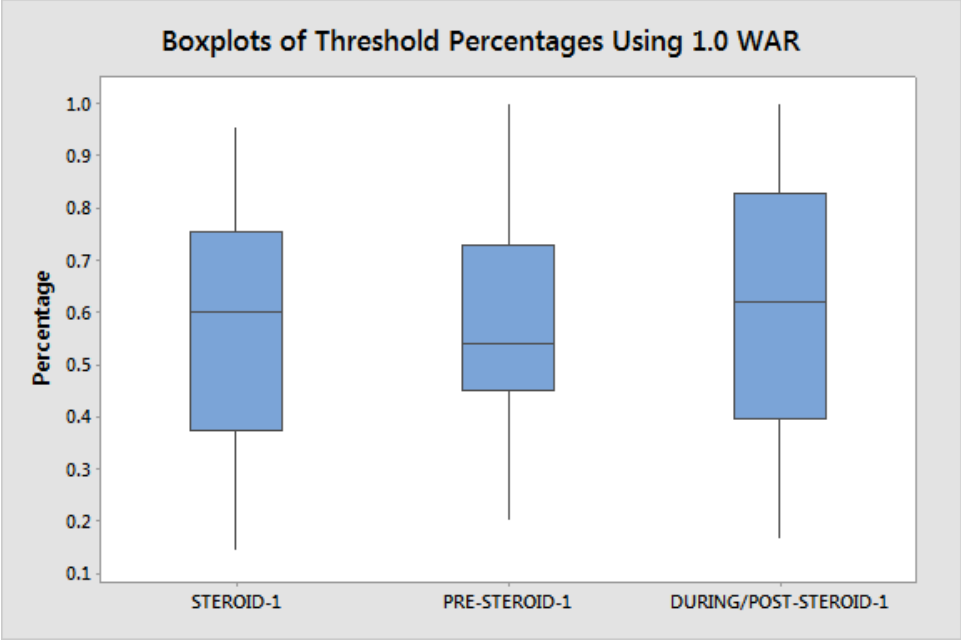

Figure 8 – Boxplots of Threshold Percentages Using WAR values of 1.0: This diagram consists of three boxplots, one for the steroid group, one for the pre-steroid era group, and one for the during/post-steroid era group. Each player's WAR scores for each season were taken into account. For instance, if a player played 10 seasons, and his WAR scores were greater than 1 for 7 out of 10 seasons, he would score a 0.7 (7/10) in this diagram. Note that the interquartile values (IQR) of the during/post-steroid group are significantly higher than the IQR values of the other groups.

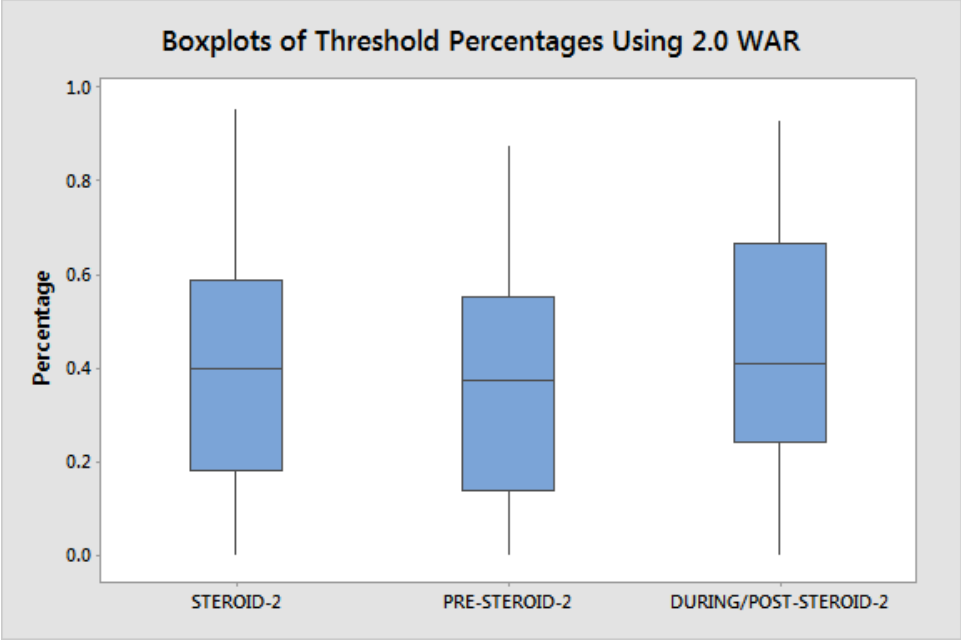

Figure 9 – Boxplots of Threshold Percentages Using WAR values of 2.0: This diagram consists of three boxplots, one for the steroid group, one for the pre-steroid era group, and one for the during/post-steroid era group. See the label for Figure 8 for a detailed explanation of the figure. Note that the interquartile values (IQR) of the during/post-steroid group are once again significantly higher than the IQR values of the other groups.

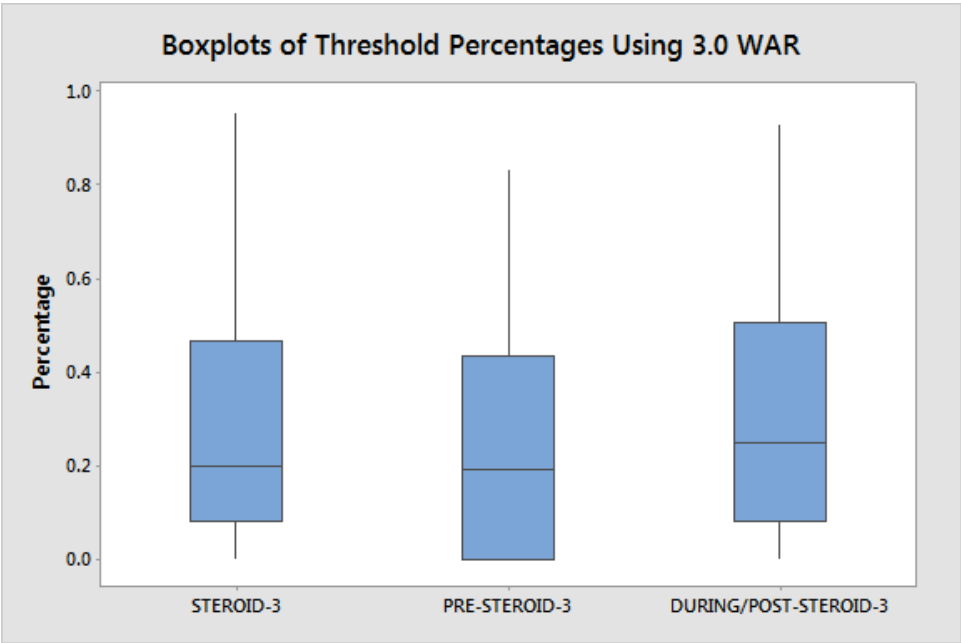

Figure 10 – Boxplots of Threshold Percentages Using WAR values of 3.0: This diagram consists of three boxplots, one for the steroid group, one for the pre-steroid era group, and one for the during/post-steroid era group. See the label for Figure 8 for a detailed explanation of the figure. Note that the interquartile values (IQR) of the during/post-steroid group are once again higher than the IQR values of the other groups.

Figure 7 is a visual representation of the percentage of values (x-axis) exceeding the WAR threshold (y-axis). So in other words, at a WAR threshold of 12, 0 percent of the values in each group exceed that threshold. Note that the during/post-steroid era group converges to zero percent the fastest whereas the steroid group converges the slowest. This is because the maximum WAR values of the during/post-steroid era group were the smallest, and the maximum WAR values for the steroid group were the largest. Although the during/post-steroid era group converges the fastest, it also appears to have the highest percentages for lower WAR values. This threshold analysis of the data is of a larger scope, so this pattern is somewhat subtle. However, this pattern becomes clearer when taking a slightly different approach in analyzing WAR thresholds for the three groups. Figures 6-8 are boxplots which better illustrate these threshold differences.

For Figures 8-10, each player's career was analyzed with respect to the threshold. For Figure 8, which utilizes a threshold value of 1.0, if an MLB player played 10 seasons and 7 of which exceeded the 1.0 threshold, his percentage would be 70 percent, or 7/10. All of these data were then used to create boxplots for the three groups using WAR threshold values of 1.0, 2.0, and 3.0. In each figure, we see that the interquartile ranges (IQRs) for the during/post-steroid era group are higher than in the other groups. Unfortunately, due to the lower percentages exceeding the WAR threshold as WAR increases, boxplots of this nature could not be produced with higher thresholds without having outliers in the data. Figure 11 represents interquartile WAR values of players in each group. Figure 12 represents ranges of players in each group (see Figures 11 & 12).

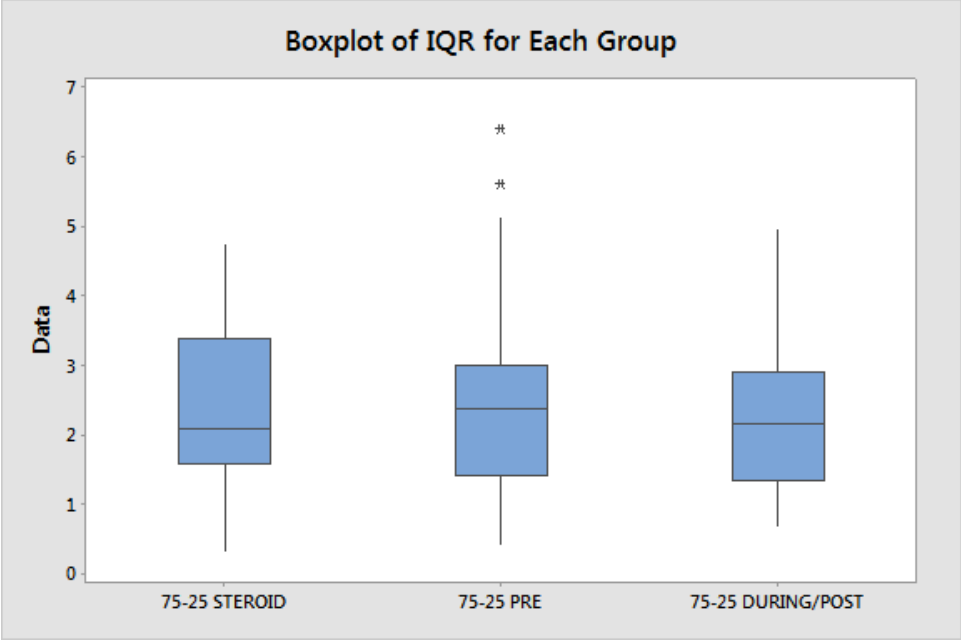

Figure 11 – Boxplots of IQR for Each Group: This diagram consists of three boxplots, one for the steroid group, one for the pre-steroid era group, and one for the during/post-steroid era group. The interquartile values (the 25 and 75 percentiles) of each MLB player from each group was calculated and then compiled into this figure.

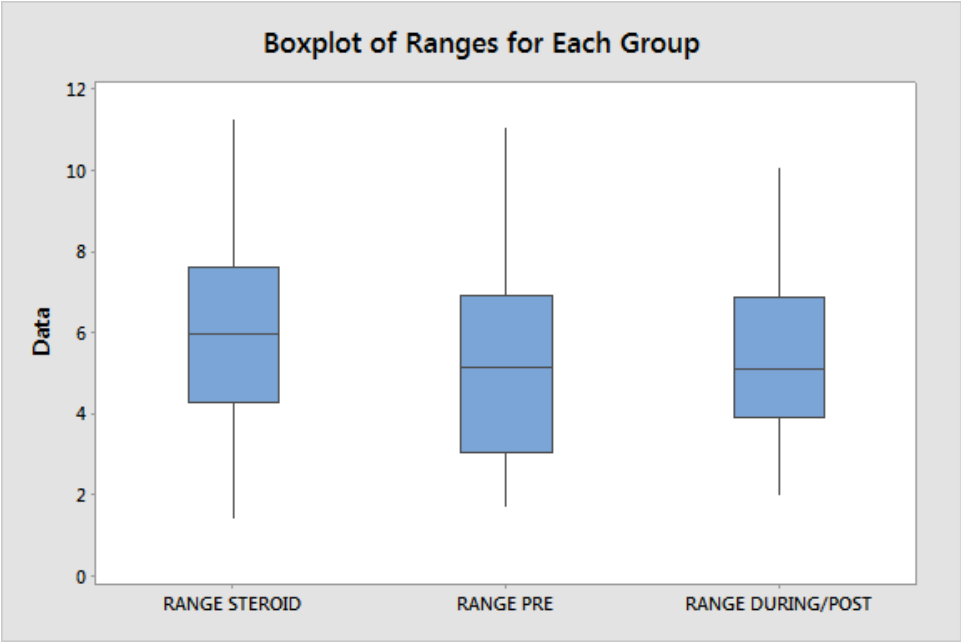

Figure 12 – Boxplots of Ranges for Each Group: This diagram consists of three boxplots, one for the steroid group, one for the pre-steroid era group, and one for the during/post-steroid era group. The ranges (maxWAR - minWAR) of each MLB player from each group were calculated and then compiled into this figure.

The two outliers in the pre-steroid era group make Figure 11 difficult to fully interpret. Regardless, the pre-steroid era group has the highest median value, which coincides with the hypothesis that this group contains the most even distribution of WAR values. Figure 12 shows that the interquartile range of the players in the steroid group is the highest, which coincides with the hypothesis that the steroid group contains the least even distribution of WAR values.

Discussion

I had initially planned on solely implementing ANOVA analysis on the fit regressions of age vs. WAR for my results section; however, the regressions of the three equations appeared to be nearly identical. Therefore, I implemented all sorts of other statistics for the analysis of the data, which can be seen in the figures in the results section. While my results overall support the hypothesis that there is more variation of WAR values in the steroid group, my results only did so at a small scale. There was nothing of high significance that supported my hypothesis of most even distribution of WAR values occurring in the pre-steroid era group and the least even distribution of WAR values in the steroid group.

However, the results of the different threshold analyses were highly suggestive. While only the steroid group and the pre-steroid era group included WAR values that exceeded thresholds higher than 10.0, it was actually the during/post-steroid era group that had the highest percentage of WAR values that exceeded lower thresholds. In other words, the during/post-steroid era group maintained the most consistency for lower to

moderate WAR values. These results make sense when considering that the value-based similarity score generator generates top matches at a career-wide level.

To illustrate this phenomenon, let's assume we have two players. Player #1 is an athlete who used steroids, whereas player #2 is player #1's most similar athlete based on the value-based similarity score generator. Assume that player #2 played after the steroid era of baseball and did not use steroids or any illegal substance. Also assume that both player #1 and player #2 played a total of three seasons and each had a cumulative total of 18 WAR in their careers. Player #1, whose WAR values have higher variation, had WAR values of 1, 11, and 6. Since the ceiling for potential value in terms of WAR is lowered due to the removal of steroids from baseball, player #2's seasonal WAR values are not going to exceed as high of thresholds, as can be seen in the results. Therefore, it would be more likely that the WAR values he puts up are say, 4, 8, and 6.

In this example, it can be seen that the steroid-free player has to maintain a higher consistency in order to have the same value at a career-wide scope that the steroid-using player has. This same concept appears to apply to the data gathered for this thesis in general. Although steroid usage is still a problem in modern MLB baseball, the data in this thesis suggest that the ban of illegal substances has been highly statistically impactful. Furthermore, because the pre-steroid era group consists of WAR values exceeding similar thresholds as the steroid group, the data suggest that steroid usage has been a part of baseball for a long time and can be traced farther back than the steroid era. This is merely a hypothesis derived from this data. It is worth noting that some of the players in the during/post steroid era group are still active (i.e. have not retired yet), so maybe this effect

is observable in data now, but will become less apparent over time. Whether or not this is the case, the overall data acquired from the various threshold analyses is telling.

As far as future research goes, it would be interesting to further analyze the alternative similarity score generator used for this thesis, as well as to perform further research on which statistics seem to be the most inflated due to steroid usage. Perhaps yet another alternative similarity score generator could be created that takes WAR into consideration to maintain an unbiased measure of player value, as well as utilizes statistics that are most heavily impacted by steroid usage. If such a similarity score generator were to exist, then this similarity score generator could potentially be utilized to produce even more telling anomalies than the ones in this study.

Works Cited

Baseball Reference. (n.d.). Retrieved October 29, 2014, from http://www.baseball-reference.com/

Boss, T. (2012, January 24). Nationals Arm Race. Retrieved February 16, 2015, from http://www.nationalsarmrace.com/?p=3743

Carise, D. (2013, September 12). Baseball and Steroids: What's the Big Deal? Retrieved November 26, 2014, from http://www.huffingtonpost.com/deni-carise/baseball-and-steroids- wha_b_3887380.html

Darowski, A. (2012, July 27). Wins Above Replacement (WAR) versus Wins Above Average (WAA). Retrieved December 2, 2014, from http://www.highheatstats.com/2012/07/wins-above-replacement-war-vs-wins-above-average-waa/

Darowski, A., Chupp, J., & Berkowitz, M. (n.d.). Hall of Stats. Retrieved November 23, 2014, from http://www.hallofstats.com/

Furnald, N. A. (2012). *The impact of age on baseball players' performance: How was this altered during the steroid era*. (Thesis, Colgate University). Retrieved from http://www.colgate.edu/portaldata/imagegallerywww/21c0d002-4098-4995-941f-9ae8013632ee/ImageGallery/2012/the-impact-of-age-on-baseball-players-performance.pdf

Grossman, M., Kimsey, T., Moreen, J., & Owings, M. (2012). *Steroids and Major League Baseball* (Berkeley University). Retrieved from http://faculty.haas.berkeley.edu/rjmorgan/mba211/steroids%20and%20major%20league%20baseball.pdf

Lenhardt, M. (2010, March 14). Illinois Business Law Journal. Retrieved February 18, 2015, from http://www.law.illinois.edu/bljournal/post/2010/03/14/the-business-of-steroids-in-baseball.aspx

Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. New York: W.W. Norton.

McGwire comes clean, admits steroids use. (2010, January 12). Retrieved December 3, 2014, from http://sports.espn.go.com/mlb/news/story?id=4816607

Meerschaert, M. (2013). Stochastic Models. In *Mathematical modeling* (Fourth ed., p. 273). Boston: Academic Press.

Mitchell, G (2007). The Mitchell Report. Retrieved April 23, 2015 from http://assets.espn.go.com/media/pdf/071213/mitchell_report.pdf

Petersen, A., Jung, W., & Eugene Stanley, H. (2008). On The Distribution Of Career Longevity And The Evolution Of Home-run Prowess In Professional Baseball. *EPL (Europhysics Letters), 83*(5), 50010-50010.

Rader, B. (1992). *Baseball: A history of America's game*. Urbana: University of Illinois Press.

Rymer, Z. (2013, January 15). How PED Testing Has Impacted Offensive Stats. Retrieved December 2, 2014, from http://bleacherreport.com/articles/1486347-analyzing-how-ped-testing-has-impacted-offensive-stats

SABR. (n.d.). Retrieved December 2, 2014, from http://sabr.org/sabermetrics

Singer, T. (2009, April 11). Report: Injuries factor in steroid use. Retrieved December 3, 2014, from http://m.mlb.com/news/article/4223696/

The Steroids Era. (n.d.). Retrieved December 1, 2014, from http://espn.go.com/mlb/topics/_/page/the-steroids-era

Vinton, N. (2014, March 28). MLB adds better drug tests, longer bans and postseason ineligibility. Retrieved February 16, 2015, from http://www.nydailynews.com/sports/i-team/mlb-adds-better-drug-tests-longer-bans-postseason-ineligibility-article-1.1738443

Waters, Z. (2008, November 21). Season similarity scores. Retrieved November 25, 2014, from http://www.hardballtimes.com/season-similarity-scores/

What is WAR? | FanGraphs Sabermetrics Library. (n.d.). Retrieved December 1, 2014, from http://www.fangraphs.com/library/misc/war/

Appendix A – Steroid Group Data

	STEROID GROUP DATA																											
	NAME	AGE	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
		WAR																										
1	Alex Rodriguez		-0.2	-0.4	9.4	5.6	8.5	4.7	10.4	8.3	8.8	8.4	7.6	9.4	4.5	9.4	6.8	4.1	4.1	3.9	2.2	0.5						
2	Alex Sanchez								-0.1	1.2	-0.1	0.2	-0.2															
3	Barry Bonds					3.5	5.8	6.2	8	9.7	7.9	9	9.9	6.2	7.5	9.6	8.2	8.1	3.8	7.7	11.9	11.8	9.2	10.6	0.6	4	3.4	
4	Benito Santiago					-0.1	3.4	3.1	2.3	2.1	2.2	1.2	0.7	1.4	1.6	2.9	0.8	0	0.8	-0.2	0.4	2.6	1.5	0.8	-0.1			
5	Brian Roberts							-0.3	0.2	2.7	2.4	7.2	3.3	4.2	5.2	2.9	1.2	0.1	-1.1	0.8	1.5							
6	Cameron Maybin				-0.4	0.8	-0.1	1.2	4.5	3	-0.8	0.3																
7	Carlos Ruiz											0	2	0.1	2.6	4	2.8	4.5	1.6	3.1								
8	Chris Davis						1	-0.9	-0.5	-0.1	1.9	6.1	1.8															
9	Chuck Knoblauch						2.8	5.3	3.6	4	6.7	8.6	6.7	2.8	3.5	0.3	0.9	-0.7										
10	David Segui							-1	-1.4	-1	1.1	-0.9	1.4	1.1	2.1	3.2	0.8	2.6	2.1	-0.1	0.3	0.1						
11	Everth Cabrera						1.4	0.1	-0.1	1.8	2.6	0.6																
12	F.P. Santangelo											0.7	3.2	1.6	1	0.8	-0.8	0.2										
13	Fernando Vina								0.1	0	0.2	1.7	1.1	3	-0.1	3.2	2.5	0.5	-0.6	0.5								
14	Gary Matthews, Jr.								0.3	-0.2	0.6	2.4	1.5	2	3	5.2	1	-1	-0.1	-0.6								
15	Gary Sheffield			-0.1	-0.2	3	-1.3	6.2	-0.2	2.1	2.2	5.9	2.6	3.4	3.2	6.1	4.4	4.4	6.8	4.2	4.1	0.4	3	-0.2	0.3			
16	Gregg Zaun								0.6	0	1	-1.1	0.5	0.7	1.1	-1.1	-0.3	2.3	3.6	1.8	2.4	0.8	1.1	0.5				
17	Hal Morris							-0.3	0	1.8	3.8	0.4	1.2	2.2	1	2.7	-0.5	0.4	0	0.7								
18	Jason Giambi								0.9	1.1	1.8	2.6	5.9	7.7	9.2	7.1	4.8	-0.1	4.6	2.8	0.9	1.9	-0.1	-0.2	1.1	-0.2	-0.3	-0.5
19	Jay Gibbons								0.9	1.6	1.5	0.5	2.2	0.4	-1.1			-0.5	-0.1									
20	Jeremy Giambi							0	0	0.2	1	2.5	-0.3															
21	Jhonny Peralta					0.6	-0.1	5.1	0.9	2.7	3.6	1	2.6	3.7	1.1	3.3	5.8											
22	Jose Canseco				0.4	3	1.6	7.3	2.2	5.4	5.3	2.3	0.1	3	2.6	3	0.2	1.5	2.7	1.1	0.8							
23	Jose Guillen					-3.3	-0.1	-1.7	0.8	0.7	-1.4	4.9	3.1	3.6	-0.9	3.5	-0.5	-2	-0.4									
24	Kevin Young							0.1	-0.3	-0.6	-0.5	-0.5	2.8	1.3	5.6	-1.5	-0.6	0.4	-0.4									
25	Lenny Dykstra						1.9	4.7	3.8	3.5	2.4	8.9	3.1	2.4	6.5	2.3	0.8	2										
26	Manny Ramirez					-0.8	1.6	2.9	4.2	4.6	5.2	7.3	4.8	5.2	6	5.4	4.1	4.4	4.5	1.1	6	2.2	0.8	-0.3				
27	Mark Carreon							-0.1	0.4	1.3	1.2	-1.3	-1.2	0.7	0.3	2.9	0.5											
28	Mark McGwire						-0.2	5.1	2.7	3.6	5.7	1.5	6.4	1.5	1.5	5.5	6.4	5.1	7.5	5.2	4.2	0.3						
29	Marlon Byrd								-0.2	3.4	-2.1	-0.1	1.2	3.2	3.5	2.2	3.7	1.9	-0.4	5.1	2.6							
30	Marvin Benard								0.3	2	-0.2	1.9	1.4	2.1	1	0.5	-0.4											
31	Matt Lawton							0.7	0.6	1.1	3.9	0.5	2.3	2.5	1.3	1.2	1.3	-0.1	-0.2									
32	Matt Williams					0	1	2.7	5	5.5	1.8	5.8	4.7	4.6	2.9	4.2	2.6	4.1	0.3	1	0.8	-0.5						
33	Melky Cabrera				-0.5	3	1.6	0.4	1.2	-0.3	4.4	4.7	-0.3	3.1														
34	Miguel Tejada							-0.2	-0.2	3.6	4.2	4.2	5.6	4.9	7.4	5.9	4.5	2.3	1.9	1.9	0.6	0		0.7				
35	Mike Cameron						0.2	-0.1	4.4	1.5	5.5	4	5.9	3.6	4.8	1.5	1.8	4.3	3.2	3.1	3	0.1	-0.2					
36	Mike Lansing									2.6	1.5	1.2	1.6	3.8	0.3	0	-1.4	0.1										
37	Mike Morse							-0.5	0.2	0.3	0	0.4	1.3	3.4	0.7	-1.5	1											
38	Mo Vaughn							-0.3	-0.2	3.2	2.7	4.3	5.6	3.9	5.6	1.8	1.8		-0.5	-0.8								
39	Neifi Perez							-0.6	1	-1.1	-0.9	1.1	1.1	-2.1	1.1	0.2	2.4	0.5	-0.2									
40	Nelson Cruz								0	0.2	-0.4	1.5	2.3	4.1	1.4	0.7	2.2	4.7										
41	Paul Lo Duca										0.1	-0.2	0.4	4.6	3.1	3.8	3.2	1.2	2.4	0.2	-0.8							
42	Rafael Palmeiro					-0.1	0.7	2.3	2.6	4.3	5.7	4.1	6.9	4	5.5	4.4	3.2	6.3	5.2	2.9	4.7	4.5	3.5	0.8	0.3			
43	Randy Velarde								-0.2	-0.3	1.9	1.4	0.3	1.8	2	1.1	3.4	2	0	0.5	7	2.3	1.9	-0.2				
44	Rondell White					0.3	0.4	3.8	1.6	4.8	3.5	2.9	2.5	2.1	1.4	2.9	2	1.7	-1.1	-0.4								
45	Ryan Braun							2	4.5	6.2	5.7	7.8	7	1.9	1.0-													
46	Sammy Sosa				-0.4	0.5	0.7	0.6	4	3.8	5.3	5.4	2.5	6.4	4.8	5.7	10.3	5.7	2.7	1.3	-1		0.1					
47	Todd Hundley					-0.4	-0.4	-0.7	-0.6	1.3	2.6	4.9	3.8	-1.3	-0.1	2.5	-1.1	0.4	0									
48	Todd Pratt									-0.1	0.5	-0.1	-0.7		0.9	0.3	0.7	1.3	-0.5	1.3	1	0.3	0.2	-0.8				
49	Troy Glaus					-0.6	3.1	7.8	5.2	4.3	1.3	1.3	3.4	4.3	3.3	4.4	-0.2	0.2										
50	Wally Joyner								3.2	4.1	3.2	2.7	2	3.8	1.3	2.7	1.4	2.2	2.8	3.7	2	0.5	0.2	-0.1				

Appendix B – Pre-Steroid Era Group Data

	PRE-STEROID ERA GROUP DATA																										
	NAME	AGE	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
		WAR																										
1	Eddie Collins			-0.1	0.1	1.4	9.7	10.5	6.5	8.8	9	9.1	9.4	7	5	3.1	5.1	7.9	4.5	3.9	6.3	5.2	5.1	4.1	2.3	0.2	-0.1	0
2	Rob Andrews						-0.6	1.7	-0.2	-0.2	0.2																	
3	Ty Cobb		0.2	2.5	6.8	6.1	9.8	10.5	10.7	9.2	7.4	5.6	9.5	8	11.3	6.6	5.5	3.3	6.7	6.7	5.5	5.4	5.8	1.7	4.4	1.9		
4	Leo Cardenas					-0.2	1.1	4.3	1.6	1.7	4.2	1.9	0.2	0	5.1	2.9	3.1	1.3	-1.2	0.6	0.7							
5	Harry Steinfeldt				1.1	0.4	0.8	1.1	3	3.9	0.1	1.3	7	4.6	1.9	4	1.9	-0.2										
6	Earl Grace						0.1		0.6	1.6	1.8	0.6	1.1	0.7	0.6	1.1	0.7	0.6										
7	John Stearns						0.1	0.2	1.3	3.4	5.3	1.5	2.3	1.9	3.8	0	-0.1											
8	Pat Mullin						-0.1	1.3					-0.7	2.4	3.5	1.2	-0.3	1.4	0.8	0.1								
9	Del Pratt								3.6	3.6	4.6	4.7	4.8	1.5	3.2	5.3	4.6	3.8	3	1.7	1.3							
10	Bruce Campbell				0.2	0.3	0.6	1.4	0.3	2	1.4	2.3	1.7	1.9	0.9	1.8	0.4											
11	Specs Toporcer						-0.1	2.2	0.7	1.3	1.8	0.1	0.1	-0.3														
12	Thurman Tucker						-0.4	1.5	3.3		2	0.2	1.1	-0.1	-0.8	0												
13	Bobby Adams								2.6	0.9	1.2	-0.1	2	0.6	2.1	1.8	1.2	1.8	0.2	0.1	0.1	-0.1						
14	Bobby Adams								2.6	0.9	1.2	-0.1	2	0.6	2.1	1.8	1.2	1.8	0.2	0.1	0.1	-0.1						
15	Harmon Killebrew		0	-0.1	-0.5	0.2	-0.1	4.2	3.1	5.7	2.8	4.2	4.7	4.3	5.9	6.5	2.5	6.2	4.9	2.8	3.1	0.5	-0.2	-0.1				
16	Milt May			0.1	1.2	1	0.6	3	0.7	0	1.2	1.4	0.5	1.6	1.4	2.8	1.6	-0.3										
17	Chuck Hinton											1.2	2.3	3.2	2.9	2.2	0.4	-0.2	0.8	0.2	1	-0.3						
18	Willie Stargell						0.1	-0.3	1.3	3.3	4.8	2.3	1.8	5.2	2.5	7.9	3.9	7.2	5.4	3.6	0.9	0.8	3.1	2.5	1.2	-0.2	0.2	
19	Farmer Weaver							-0.4	0.7	1.4	1.7	0.8	1.5	-0.9														
20	Don Lenhardt											1.6	1.1	0.6	1.1	-0.6												
21	Roy Smalley						1.1	3	1.6	5.9	4.3	3.4	0.6	2.8	2.8	0.2	1	0.9	0.2									
22	Ken Singleton							1	1.6	2.1	5.7	1.5	5	3.4	5.7	4.7	5.3	4	1.2	0.6	1.9	-1.8						
23	Deron Johnson					0	-1.5	-0.3		3.1	4	-0.1	-0.5	-0.7	0.5	1.4	3.2	-1.2	1.2	-2.2	-0.2	-0.5						
24	Cliff Carroll						-0.3	0.5	2.1	1.2	0.5	-0.3	-0.5		2.5	1.2	1.9	-0.2										
25	Hughie Jennings						0.9	1.8	-0.9	4.8	7.5	8.3	7.3	7.5	1.9	0.9	1.2	1.3	0									
26	Willie McCovey					3.1	1.7	2.6	1.9	6.5	-0.1	5.9	5.9	5	7	8.1	6.6	2.6	-0.5	3.1	3.2	1.6	0.3	1.3	-0.7	0.2	-0.7	
27	Vin Campbell				0		1.7	0.2	0.3		1	1.2																
28	Harry Heilmann			-0.3		2.6	2.7	2.1	3.8	2	6.8	5.4	9.3	6.4	6.9	5.2	7.2	4	3.7	4.6		-0.1						
29	Bill North							0.1	-0.6	7	4.2	5.7	2.6	0.1	2.2	2.5	2.3	0.7										
30	Gary Geiger					0.9	1.2	1.3	1.7	1.6	2	0.2	0.4	0.3	-0.6		0	-0.1										
31	Bruce Campbell				0.2	0.3	0.6	1.4	0.3	2	1.4	2.3	1.7	1.9	0.9	1.8	0.4											
32	Jimmy Collins									1.7	2	5.3	6.9	3.9	3.1	6.7	4.3	5.3	5.3	4.6	0.9	3.1	0					
33	Ira Flagstead							-0.1		3.1	0.5	0.2	0.8	2.4	2.9	1.8	2.4	2.5	2.6	-0.4	-0.6							
34	Ed McKean							0.9	4	4.5	5.3	3.1	1.2	3.3	3.7	4.4	3.7	0.7	3.4	0.3								
35	Tommy Leach				-0.1	1.3	0	3	6	3.5	3.7	1.7	2.6	5.4	4.8	3.5	1.3	0.4	1.9	4	3.6	0.4	0.1					
36	Bill Kuehne								0.4	3	0.7	0.1	2.3	1.9	0.6	-0.1	0	-1.9										
37	Fred Mann								-0.4	0.6	2.7	0.7	1.6	1.5														
38	Bob Watson				0	0	-0.4	0.4	0.1	1	2.9	4.7	0.7	3.3	3	4.7	3	2.9	2.4	-0.1	0	0.3	-0.3					
39	Ski Melillo										-0.2	-2.2	-0.9	3	-0.9	1.8	-1.8	-0.2	-1.9	-0.4								
40	Nate Colbert				-0.2		-1	2.6	2.6	3.7	5.4	2.8	0.2	-1.8	-0.1													
41	Ernie Whitt								0.1	-0.2	0.1		0.1	0.2	0.9	3.4	1.7	2.5	2.6	2.6	2.8	2.7	-1.2	0.2				
42	Dwight Evans				0.4	0.9	4.5	5.1	3.8	2	3.1	3.7	3	6.7	6.4	1.2	5.4	4.1	4.4	4.8	3.2	3	0.9	0.6				
43	Woodie Held						0			2.9	-0.5	3.5	3.8	3.9	3.8	2.8	1.2	0.7	0	1.1	-0.7	-0.5						
44	Amos Strunk			0	-0.4	0.6	0.6	4.1	2.3	3	4.1	5.6	3.6	1.8	0.1	0.5	1.8	0.4	0.2	-0.4								
45	Gavvy Cravath											1.6	0.3			2	5.6	3.9	7	3.5	4.4	0.7	3.4	0.4				
46	Zack Wheat					0.7	2.7	1.7	2.9	2.9	4.9	2.7	6	2.9	2.7	2.7	4.7	2.7	4.3	2.5	6.7	5	1	0.5				
47	Eddie Bressoud								-0.2	0.4	-0.2	2.8	1.8	-0.6	4.8	2.8	4.6	-0.3	2.1	-0.9								
48	Bob Fisher									-0.8	0.3	0.7	3.1	0.4		3	0											
49	Harlond Clift					1.2	3.2	5.3	6.7	6.1	3.2	4.4	3	3.8	1.1	-0.4	1.9											
50	George Burns					-0.3	0	2.1	6.4	1.8	3.2	6	4.3	5.1	4.6	2.9	1.3	2	-0.1	-0.1								

Appendix C – During/Post-Steroid Era Group Data

	DURNG/POST-STEROID ERA GROUP DATA																												
		AGE	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	
1		WAR																											
2	Mike Schmidt						0.1	1.9	9.7	7.7	8	8.8	6.2	7.9	8.8	7.7	7.4	6.8	7	5	6.1	6.1	1.8	-0.4					
3	Dave Berg											1.2	1.6	0.2	0	0.5	-1.1	-1.2											
4	Albert Pujols					6.6	5.5	8.6	8.5	8.4	8.5	8.7	9.2	9.7	7.5	5.3	4.8	1.9	3.9										
5	Chris Speier					2.7	6	1.1	2.7	4.2	-0.2	1.6	3.1	1	1.4	-0.1	1	1.4	0.1	-0.2	1.2	2.2	1.4	0					
6	Lloyd Moseby				-0.7	0.4	0.6	6	7.3	3.1	2	4	1.9	1.4	1	0.5													
7	Ryan Freel									0		0.5	3.8	2.6	2.7	-0.2	-0.3	-0.4											
8	Yunel Escobar								2.4	3.4	4.3	2.3	4.7	2.9	3.3	-0.2													
9	Seth Smith								0.2	-0.3	2.3	0.4	0.5	1.8	0.6	3.9													
10	Andy Van Slyke						1.8	1.7	3.5	3.3	5.5	6.4	3.5	4.3	3.6	6	1.1	0.5	0										
11	Dmitri Young						-0.3	0.3	1.6	2	1	0.9	0.4	3.4	1.7	1.4	-0.2	0.2	-0.5										
12	Tadahito Iguchi														2.8	2.1	1.3	0.1											
13	Dustin Ackley							3.8	2.6	0.7	1.9																		
14	Mike Pagliarulo								0.8	1.6	2.8	1.6	-0.8	-1.3	0.5	2.7	-0.1	2.8	0										
15	Chris Young						0.2	0.7	1.4	0	5.4	5	2	-0.2	0.9														
16	Jim Thome				-0.1	-0.1	1.2	2	5.9	7.5	5.5	3.3	4.5	4.7	5.6	7.4	4.7	3.2	0.2	4.9	3.6	2.1	1.3	3.6	1.6	0.5			
17	Kurt Suzuki							0.4	3.8	3.4	2.2	1.6	0.2	0.1	2.2														
18	Rob Deer							0.3	0	1	2	2.5	0.2	1.1	1.1	3.9	1.3			0.5									
19	Fred McGriff						-0.1	1.4	6.2	6.6	5.2	3.4	5.2	4.1	4.5	1.4	1.7	0.2	2.9	4	0.2	3.7	2.1	0.4	-0.6				
20	Xavier Nady					0.1			1.1	0	0.4	0.4	0.6	3.5	-0.1	-1	-0.3	-0.8	-0.2										
21	Bubba Trammell									-0.5	1	1.4	0.4	2.2	-0.2	-0.1													
22	Terry Steinbach								0.2	3.5	2.4	0.7	0.7	1.7	4	2.5	3.1	2.6	3.4	0.6	1.4	1.2							
23	Moises Alou							-0.2		2.5	2.5	5.1	1.2	1.2	3.5	6.2		2.6	2.9	0.2	1.1	4	3.4	1.4	2.2	0.1			
24	Hubie Brooks							0.5	2.6	-1.4	0.2	2.7	2.5	4.7	0.2	2.5	-1.2	1.6	0.2	-1.3	-0.4	-0.7							
25	Travis Lee							1.1	-0.5	0.4	1.1	1.1	3.6	-0.3	1	-0.1													
26	Dustin Pedroia						-0.8	3.9	6.9	5.6	3.2	7.9	5.1	6.6	4.8														
27	Frank Thomas						2.3	7	7	6.2	6.3	5.3	5.5	7.3	3.5	2.3	6	0	1.9	4.3	2.8	0.4	3.2	2.2	0.2				
28	Bubba Trammell									-0.5	1	1.4	0.4	2.2	-0.2	-0.1													
29	Edgar Martinez								0.3	-0.1	0.5	5.5	6.1	6.5	0.2	3.1	7	6.5	6.2	5.6	4.9	5.7	4.8	2.6	3.3	-0.3			
30	Casey Blake									-0.1	0	0	-0.2	3.5	3.5	2	2.3	2.8	2.7	4.6	2.7	0.9							
31	Rich Becker						0	0.3	-0.7	4.3	2.7	-0.4	1.4	0.8															
32	Kevin Millar											0.1	0.4	1.7	2.8	2.5	1.1	2.8	1.1	0.4	1.3	0.5	-0.7						
33	Dustin Pedroia							-0.8	3.9	6.9	5.6	3.2	7.9	5.1	6.6	4.8													
34	Derek Bell							-0.1	0.8	-0.4	1	1.9	2.8	2.7	5.4	-1.4	1.6	-1.2											
35	Jay Bell				0.2	-0.3	-0.7	0.7	2.5	3.8	3.9	6.2	3.4	1.2	2	5.4	3.9	4.9	0.8	0.4	-0.3	-1.1							
36	Ian Kinsler							1.9	4.1	4.7	6	4	7.1	2.4	4.5	5.5													
37	Miguel Cairo						0.1	0.1	3.2	0.7	-0.2	0.8	-0.1	0	1.3	-0.5	0.8	0.2	0.3	0.3	0.7	1.6	-1.3						
38	Craig Wilson								1.1	0.3	1.3	1.2	0.5	-0.7	-0.2														
39	Ryan Klesko					-0.3	0.4	1.2	1.6	3.5	0.4	2.9	1.1	3.2	4.6	4.2	0.8	1.2	1.7	0.2	0.3								
40	Cesar Izturis					0.3	-0.9	0.2	3.8	-0.2	3.8	-0.2	0	-0.2	1.7	1.3	0.3	0.1	-0.6	0.4									
41	Corey Hart						0	-0.2	-0.4	4.7	1.3	1.1	3.9	3.4	1.9		-0.5												
42	Yunel Escobar								2.4	3.4	4.3	2.3	4.7	2.9	3.3	-0.2													
43	Eddie Murray					3.2	4.3	4.9	4.4	3.7	5.2	6.6	7.1	5.6	4.1	3.8	3.2	2	5.1	1.2	1.6	1.1	-0.1	2.4	-0.3	-1			
44	Aaron Rowand							1.4	1.3	0.8	5.6	3.7	0.5	5.1	0.6	0.9	0.4	0.5											
45	Von Hayes						0.6	2.9	-0.1	4	3	4.9	3.5	2.2	5.1	3.1	1.3	-0.8											
46	Joey Votto							0.1	3.3	4.8	6.9	6.3	5.9	6.4	1.9														
47	John Olerud				0	1.7	1.8	3.3	7.7	3.2	2.2	2.5	4.1	7.6	5.6	3.6	5.2	5.1	2.7	1.1	0.7								
48	Mike Macfarlane							-0.1	0.5	-0.2	0.7	2.7	2.2	3.1	1	1.1	2.9	0.3	0.4	0.1									
49	Tony Eusebio								-0.1			1.1	1.9	0.1	0.1	0.1	0.8	0.4	0.1										
50	Ray Lankford							0.8	1.6	4.6	2.3	2.8	3.8	5	5.2	6.2	3.7	1	1.1	-0.4	0.4								
51	Tino Martinez						-0.1	-0.3	0.3	2.2	1.1	4.5	2.1	5.1	3.2	2.2	0.3	2.2	1.4	0.9	2.3	1.5							

Appendix D – Code Used For Threshold Analysis

(compiled using Dev C++)

```
#include <fstream>
#include <iostream>
#include <cstdlib>

using namespace std;


// declaring functions
int analysis( char filename[], double threshold );
double percent( char filename[], double threshold );

int main()
   {
    // initializing variables
    char filename[100];
    char filename2[100];
    double perc = 0;
    ofstream fout;
    double counter = -3.5;

    // enter name of text file with baseball data
    cout << "Enter filename with data: ";
    cin >> filename;
    cout << endl;
    cout << "Enter filename to upload calculations to: ";
    cin >> filename2;

    // create new file for analysis
    fout.clear();
    fout.open( filename2 );

    // run through data in the text file (counter = threshold)

    while( counter < 15 )
         {
          // analyze data
          perc = percent( filename, counter );
          // read out data
          cout << perc << endl;
          // upload data to new file
          fout << perc << endl;
          // cycle through for threshold
          counter = counter + .1;
```

```
        }
        cout << endl << endl;

    // pause program
    fout.close();
    system( "pause" );

    return 0;
   }


// test function implemented in prior program (but not this one)
int analysis( char filename[], double threshold )
   {
    // variables
    ifstream fin;
    int counter = 0;
    int age;
    double WAR;

    // open file
    fin.clear();
    fin.open( filename );

    fin >> age >> WAR;

    if( threshold < WAR )
       {
        counter++;
       }

    cout << filename << endl;

    // counter
    while( fin.good() )

        {
         fin >> age >> WAR;
         if( threshold < WAR )
            {
             cout << age << "\t" << WAR << endl;
             counter++;
            }
        }
    cout << endl;
```

```
    // close file and return answer
    fin.close();
    return counter;
   }


double percent( char filename[], double threshold )
   {
    // variables
    ifstream fin;
    double counter = 0.0;
    double counter2 = 0.0;
    double percent = 0.0;
    int age;
    double WAR;

    // open file
    fin.clear();
    fin.open( filename );

    fin >> age >> WAR;

    if( threshold < WAR )
       {
        counter++;
       }
    counter2++;

    // counter
    while( fin.good() )
         {
          fin >> age >> WAR;
          if( threshold < WAR )
             {
              counter++;
             }
              counter2++;
         }

    percent = (counter / counter2) * 100.0;

    // close file and return answer
    fin.close();
    return percent;
   }
```

Appendix E – Boxplot Values

FIGURE 5 VALUES			
	STEROID	PRE	DURING/POST
Q1	1.61	1.09	1.48
Median	3.44	2.96	2.66
Q3	5.52	4.61	4.98
IOR	3.9	3.52	3.51
Low Whisker	0.34	0.24	0.43
High Whisker	8.91	8.93	9.27
N	50	50	50

FIGURE 6 VALUES			
	STEROID	PRE	DURING/POST
Q1	1.25	1.01	1.2
Median	1.62	1.56	1.54
Q3	2	1.87	1.83
IOR	0.75	0.86	0.63
Low Whisker	0.76	0.44	0.57
High Whisker	2.7	2.61	2.78
N	50	50	50

FIGURE 8 VALUES			
	STEROID	PRE	DURING/POST
Q1	0.38	0.45	0.4
Median	0.6	0.54	0.62
Q3	0.75	0.73	0.83
IOR	0.38	0.28	0.43
Low Whisker	0.14	0.2	0.17
High Whisker	0.95	1	1
N	50	50	50

FIGURE 9 VALUES			
	STEROID	PRE	DURING/POST
Q1	0.18	0.14	0.24
Median	0.4	0.37	0.41
Q3	0.59	0.55	0.67
IOR	0.41	0.41	0.42
Low Whisker	0	0	0
High Whisker	0.95	0.88	0.93
N	50	50	50

FIGURE 10 VALUES			
	STEROID	PRE	DURING/POST
Q1	0.08	0	0.08
Median	0.2	0.19	0.25
Q3	0.47	0.44	0.51
IOR	0.38	0.44	0.43
Low Whisker	0	0	0
High Whisker	0.95	0.83	0.93
N	50	50	50

FIGURE 11 VALUES			
	STEROID	PRE	DURING/POST
Q1	1.575	1.42	1.35
Median	2.08	2.38	2.15
Q3	3.37	3	2.9
IOR	1.79	1.58	1.55
Low Whisker	0.3	0.4	0.68
High Whisker	4.75	5.13	4.98
N	50	50	50

FIGURE 12 VALUES			
	STEROID	PRE	DURING/POST
Q1	4.28	3.08	3.9
Median	6	5.15	5.1
Q3	7.63	6.93	6.9
IOR	3.35	3.85	3
Low Whisker	1.4	1.7	2
High Whisker	11.3	11.1	10.1
N	50	50	50