Discriminant Function Analysis
Of Major League Baseball Steroid Use

by

Kristen Renee Poole

(Under the direction of Nicole Lazar)

Abstract

Statistics plays an important role in all areas of sports as a method of keeping track of players' performances. What also plays a role in sports are performance-enhancing drugs. Major League Baseball has made attempts to rid baseball of these illegal substances by implementing drug tests, but unfortunately the problem persists. The purpose of this thesis is to find an effective way of determining steroid use in baseball based on players' batting statistics. A comparison will be made between two groups of Major League Baseball players: those who have ever been suspended for the use of or involvement with performance-enhancing drugs and those who have never been suspended for nor suspected of performance-enhancing drug use. The results will be used to determine the status of a third group of Major League Baseball players, those who have ever been suspected of steroid use.

Index words: Major League Baseball, Sports Statistics, Steroids,
Performance-Enhancing Drugs, Discriminant Analysis, Logistic Regression

Discriminant Function Analysis
Of Major League Baseball Steroid Use

by

KRISTEN RENEE POOLE

B.S., LaGrange College, 2012

A Thesis Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Master of Science

Athens, Georgia

2014

Discriminant Function Analysis
Of Major League Baseball Steroid Use

by

Kristen Renee Poole

Approved:

Major Professor:	Nicole Lazar
Committee:	Jack Morse Jaxk Reeves

Electronic Version Approved:

Dr. Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2014

Discriminant Function Analysis

Of Major League Baseball Steroid Use

Kristen Renee Poole

April 28, 2014

Acknowledgments

First, I would like to express my sincere gratitude to my major professor, Dr. Nicole Lazar. She provided me with the support I needed to accomplish this task and without her it would not have been possible. I would also like to thank her for being a positive influence throughout the entire process.

Additionally, I would like to thank Dr. Jaxk Reeves for helping me to improve the quality of my dataset and ultimately my entire study. A special thanks goes to Jack Morse for being a part of my team as well.

My family has been there for me since the very beginning, and I could not be more thankful for their support and encouragement. My fiance has incessantly surrounded me with positive energy which helped me tremendously during times of discouragement and bumps in the road. Thank you all for each and every thought and prayer you have sent out along the way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Performance-enhancing drugs, more commonly referred to as PEDs by sports commentators and broadcasters alike, are one of Major League Baseball's biggest problems. These substances are the topic of many conversations among sports fanatics and professionals across the globe. Some even wonder if they could be detrimental to America's favorite pastime. The implementation of drug tests, although a good attempt at preventing PED use in Major League Baseball (MLB), has still not reached its ultimate goal, which is to stop PED use in the MLB altogether. A few players have even mastered the art of taking steroids yet still receiving that coveted negative test result, as revealed in Anthony “Tony" Bosch's interview with ESPN that took place in January of this year. He spoke in regards to a recent issue concerning Alex Rodriguez, the New York Yankee's highest paid player (Barry, 2014). Just like players have found ways to manipulate drug tests, perhaps it is possible for us to find other ways of determining if they are truly following MLB guidelines against steroids and other PED use. The idea of using statistics to solve this ongoing problem comes to mind since statistics plays such an integral role in all sports. In this thesis, we will use linear discriminant analysis in hopes of working toward a better solution to Major League Baseball's top-rated issue.

1.1 History of MLB's Drug Policy

On April 26, 1990, Congress introduced the Anabolic Steroids Control Act, making these drugs illegal. Anabolic steroids were defined by Congress as muscle growth promoting drugs that are similar to testosterone. Major League Baseball's commissioner, Fay Vincent, stated in 1991 that this act was to be followed by all MLB players and personnel. He prohibited not only the use of these drugs, but also the possession or sale of them. What his statement lacked, however, was how MLB planned to enforce this rule. There would be no mention of drug tests until 2001, ten years later.

In the mid to late 1990s the league had some powerhouse hitters on its hands. In 1996, seventeen players hit forty or more homeruns, and in the following season another twelve players hit the 40 home run mark or passed it. By 1998 it was apparent that this was not pure talent and therefore no coincidence. Roger Maris' single-season record of 61 homeruns was broken by both Sammy Sosa and Mark McGwire. Also, for the first time in MLB, four players hit fifty or more homeruns in one season.

Halfway through the year, it was discovered that Mark McGwire was hiding a PED in his locker. A bottle of Androstenedione was found there, but sparked very little controversy. He went on to hit a record of seventy homeruns that same year with no complaints from Major League Baseball regarding his drug use. Apparently nothing stood in his way.

By 2000, the league's slugging percentage was an all-time high of .437. The slugging percentage is a measure of the hitter's power. It is calculated by dividing the total bases (singles + doubles + triples + homeruns) by the number of times a player goes up to bat. For the most part, the league's slugging percentage has remained above .400 since the year 2000, with the exception of a .399 in 2011. By this point, MLB wanted to implement drug testing, but could not do it without the permission of the Major League Baseball Players Association (MLBPA). The association at the time was being charged for protecting those players who were thought

Table 1.1: Minor League Baseball Punishments for PED Violations in 2001

Offense	Ban
1st	15 Games
2nd	30 Games
3rd	60 Games
4th	One Year
5th	Permanent

to be taking PEDs. Since they were unable to enforce drug tests on MLB players, they decided to start somewhere smaller, the minor leagues. The penalties for these players are listed in Table 1.1.

These so-called punishments seem more like an excessive amount of warnings to most; nothing too serious. These penalties began at the beginning of 2001 for Minor League Baseball. That same year Barry Bonds broke Mark McGwire's record with seventy-three homeruns. Sammy Sosa had sixty homeruns, and fifty homeruns were hit by both Luis Gonzalez and Alex Rodriguez. Years later, the last three players mentioned would all eventually admit to having taken PEDs during this season.

Tom Verducci of *Sports Illustrated* wrote an article in 2002 that finally convinced the MLBPA that the MLB should be able to test the players for drugs. In his article, Verducci states,

> Steroid use, which a decade ago was considered a taboo violated by a few renegade sluggers, is now so rampant in baseball that even pitchers and wispy outfielders are juicing up and talking openly among themselves about it. According to players, trainers and executives interviewed by SI over the last three months, the game has become a pharmacological trade show. (bleacherreport.com)

Once this article was released to the public, the MLBPA finally reacted. The organization knew fans were beginning to question the authenticity of the players, and this article would only exacerbate the situation. Random testing began in 2003, but with no penalties. The purpose was merely to determine exactly what percentage of MLB players were actually taking

Table 1.2: MLB Punishments for Drug Violations as of 2004

Offense	Ban
1st	10 Days
2nd	30 Days
3rd	60 Days
4th	1 Year

Table 1.3: MLB Punishments for Drug Violations as of 2005

Offense	Ban
1st	50 Games
2nd	100 Games
3rd	Lifetime

steroids or other PEDs. If the percentage was less than five percent, MLB would drop the issue, but the percentage was somewhere between five and seven percent. Mandatory testing began in 2004, but again with only lax consequences. First-time offenders would have to be treated for steroid use and would not even have their names publicly released. Second-time offenders would be subject to discipline (Rymer, 2013).

In December of 2004, Commissioner Bud Selig urged the union to enforce a stricter drug policy and Chief Donald Fehr agreed (Rymer, 2013). They established several consequences that MLB players would face if they were to fail a drug test or have other evidence of drug use against them. The consequences as of December 2004 are shown in Table 1.2.

By the end of the 2005 season, twelve players had been suspended for the use of PEDs. This was a fairly high number, causing the MLB to finally realize that the punishments were still not quite harsh enough. The new penalties are shown in Table 1.3. This last set of punishments for MLB drug policy violations are still in effect to this day. While the league's power numbers have dropped considerably, the issue of PED use has not been halted just yet.

1.2 Drug Testing

Steroids and other PEDs can be injected or taken orally. The choice of method determines how long the drug stays in the player's system. Injected PEDs typically last longer than oral drugs. Oral PEDs will stay in the body for about three to four weeks while injected PEDs can stay in the body for about three to four months.

Major League Baseball's drug policy states that players will be tested upon arrival to spring training. A urine sample is taken during the time of their spring training physicals. Players are also subject to unannounced blood tests for hGH (the human growth hormone) during spring training. All players are randomly selected for another urine test at some point during the season. The second test is unannounced. Plus, 1400 additional tests of randomly selected players take place throughout the season.

When MLB enforced the new punishments in 2005 for PED use, possession, sale, or distribution, it also decided that MLB players' names would be released publicly with their first offense. The names of any MLB players who have been suspended at any point during their baseball career along with the date on which they were suspended are listed online for the public to view at www.fannation.com. These names play an important role in the data set necessary for this study.

Chapter 2

Data and Methods

2.1 Compilation of Data

Rather than using an existing data set, we compile the data for this thesis to create a new data set. The initial data set consists of players of all positions, including pitchers. However, only the batting statistics of these players are being analyzed. Therefore, the inclusion of pitchers in the data set would eventually lead to inaccurate results considering pitchers of the American League do not bat and pitchers of the National League bat less often than position players. When the National League pitchers must bat, they typically produce lower batting statistics compared to those of position players. This is because they spend the majority of their time perfecting their pitching rather than worrying about their batting.

For those readers who are unfamiliar with baseball, Major League Baseball is divided into two leagues, the American League and the National League. Each league is composed of three divisions, the East, Central, and West divisions, and each division is composed of five MLB teams, giving a grand total of thirty teams in the MLB. It may also be important to note that 37.5% of players ever suspended for PEDs have been pitchers. It is understood that they use the drugs for the purpose of ensuring a shorter healing process so that they are able to participate

Table 2.1: Variables of the Data Set

	Variable	Meaning
1	G	Games
2	AB	At Bats
3	PA	Plate Appearances
4	R	Runs
5	H	Hits
6	Double	Doubles
7	Triple	Triples
8	HR	Homeruns
9	RBI	Runs Batted In
10	SB	Stolen Bases
11	CS	Caught Stealing
12	BB	Walks
13	SO	Strikeouts
14	BA	Batting Average
15	OBP	On-Base Percentage
16	SLG	Slugging Percentage
17	OPS	On-Base Percentage + Slugging Percentage
18	OPSplus	OPS Adjusted to Player's Ballpark
19	TB	Total Bases
20	GDP	Double Plays Grounded Into
21	HBP	Hit By Pitch
22	SH	Sacrifice Hits
23	SF	Sacrifice Flies
24	IBB	Intentional Walks
25	Name	Name of Player
26	POS	Position

more frequently and pitch during consecutive games.

The final data set now consists of three groups or classes of MLB players. The first group includes only players who have ever been suspended from Major League Baseball at any time for PED-related reasons. The list of suspended players is gathered from two webpage articles, Baseball's Steroid Era and FanNation. This first group consists of twenty-four players. The second group consists of players who have ever been suspected of using steroids due to reasons including, but not limited to, above-average performance as discussed by ESPN or other sports networks. The list of thirty-four players suspected to have used steroids but never tested positive or had other definitive evidence against them, can be found at www.fannation.com and www.about.com. The third and final group consists of twenty-four players who have never been suspended from baseball for PED involvement and have never given reason to believe that they have used steroids or other PEDs.

These non-steroid users are matched with the twenty-four steroid users to yield a total sample size of $24+34+24=82$ players. For instance, say the first player of group one was an outfielder that played for twelve years, from 1991 to 2002. Then player one of group three is also an outfielder that played for twelve years, from 1991 to 2002. In the case that there does not exist another outfielder who matches this exact description, another player is selected who played during the same time period, with the exception of one or two years. So we might have a player matched to player one of group one who played from 1990 to 2001 or 1992 to 2002. To ensure that the players are selected at random, the thirty MLB teams are put into a drawing. One by one, a team is selected and that team's all-time roster is inspected for a matching player until the data set consists of twenty-four non-steroid users. These all-time rosters can be easily found on each team's personal webpage.

A new variable, IS, is created to indicate which of the three groups each player belongs to. $IS=1$ indicates a suspended player, $IS=0.5$ indicates a player who has ever been suspected of PED use, and $IS=0$ indicates a player who has never been suspended for nor suspected of PED use. The rest of the variables included in the final data set can be referenced in Table 2.1. Most of these variables are self-explanatory, however some are not. For instance, GDP, or double plays grounded into, is the number of times the player hits a ball on the ground and causes two outs in the process. This usually means that an out is made at second base and then first base, but that is not always the case. Total bases= $Single+2*Double+3*Triple+4*HR$, where single is the number of times he hit the ball and made it to first base, double is the number of times he hit the ball and made it to second base, triple is the number of times he made it to third base, and HR, or homeruns, is the number of times he hit the ball and made it all the way to home base. The slugging percentage is the total bases per at bats. On-base percentage is the sum of his hits, walks, and number of times he was hit by a pitch, divided by the sum of his at bats, walks, sacrifice flies, and number of times he was hit by a pitch.

At this point, the data set consists of a total of eighty-two players, each with anywhere from one to twenty lines of data per player, depending on how many years he played in Major League Baseball. To reduce the data set to eighty-two lines of data, one line of data per player, the following formulas are used.

The variables that are averages, such as Batting Average, On-Base Percentage, etc., are averaged in a slightly different way than the other variables, accounting for the number of plate appearances of each individual player. It is possible that there exists one player in the dataset who only played for one year while there is another player who played for twenty years. Therefore, rather than averaging the sum of each batting statistic over the number of years of each player's baseball career, the batting statistic for each individual year of each player is first multiplied by his number of plate appearances for that same year. The sum of $PA_i * Y_i$ is taken for each player and is then divided by his total number of plate appearances. By doing this, plate appearances are taken into consideration, and we are able to control for the differences in the players' career lengths. The following formula may give a better perspective.

$$\frac{\sum PA_i * Y_i}{\sum PA_i} \tag{2.1}$$

for i=l,2,3,...,k where k=the length of the player's baseball career in years and Y_i are variables that are averages. The remaining variables, such as Games, Plate Appearances, Runs Batted In, etc., are simply averaged over the number of years of each player's baseball career.

2.2 A Visual Comparison

Now that the dataset is compiled and completely organized, we begin our study with some exploratory data analysis in order to examine basic trends in the data. The scatterplots in Figures 2.1, 2.2, and 2.3 show the comparison between the steroid users and non-steroid users for slugging percentage, runs batted in, and homeruns, respectively. These are some of the variables that are expected to be good predictors of steroid use prior to analysis.

The points are plotted in pairs, each pair containing a value for a user and a value for the non-user who is matched with him by career length and position. Steroid users typically have higher offensive statistics than non-steroid users. According to Figure 2.1, for seventeen out of the twenty-four pairs, the steroid user has a higher slugging percentage than his paired non-user. The pairs in Figures 2.2 and 2.3 show that most steroid users also have more homeruns and runs batted in than their paired non-users.

Based on the figures, the two groups obviously differ. The question is to what extent do they differ and which variables are the best at distinguishing between the two groups. Since we wish to determine the best separating factors of the two groups, discriminant function analysis seems to be a logical approach. It will show how much the groups of steroid and non-steroid users contrast and will output a function that separates the two groups of data.

2.3 Logistic Regression and Linear Discriminant Analysis

Discriminant analysis is a method that is widely used for multivariate data. The techniques used for this method aim to determine the distinguishing factors between two or more groups. The most common technique for the analysis of two groups is known as Fisher's Linear Discriminant Function. Once the function is developed, it can be used to assign the individuals of a third group to their correct group based on the function and using the Classification Rule, which is

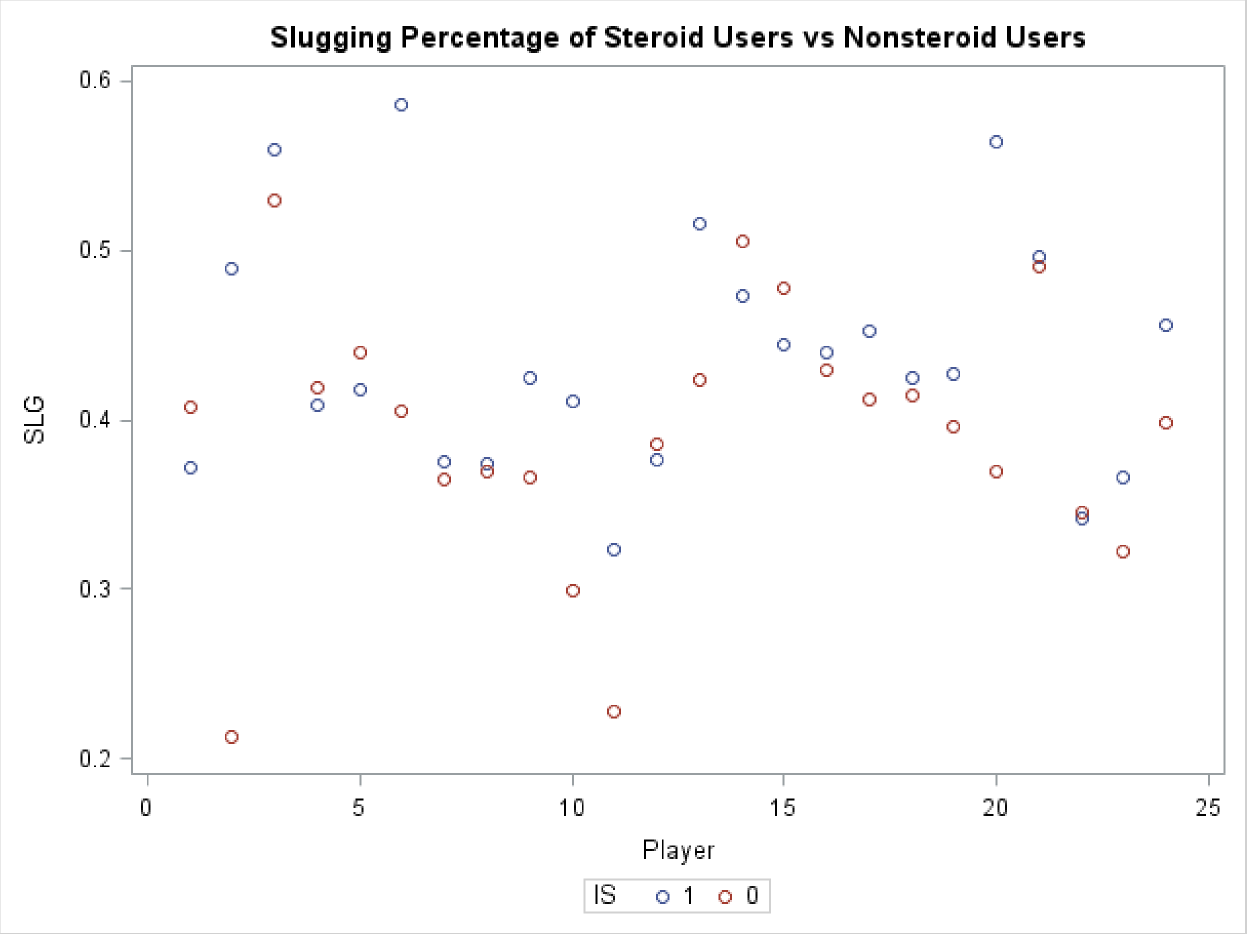

Figure 2.1: Slugging Percentages of Steroid Users vs Nonsteroid Users

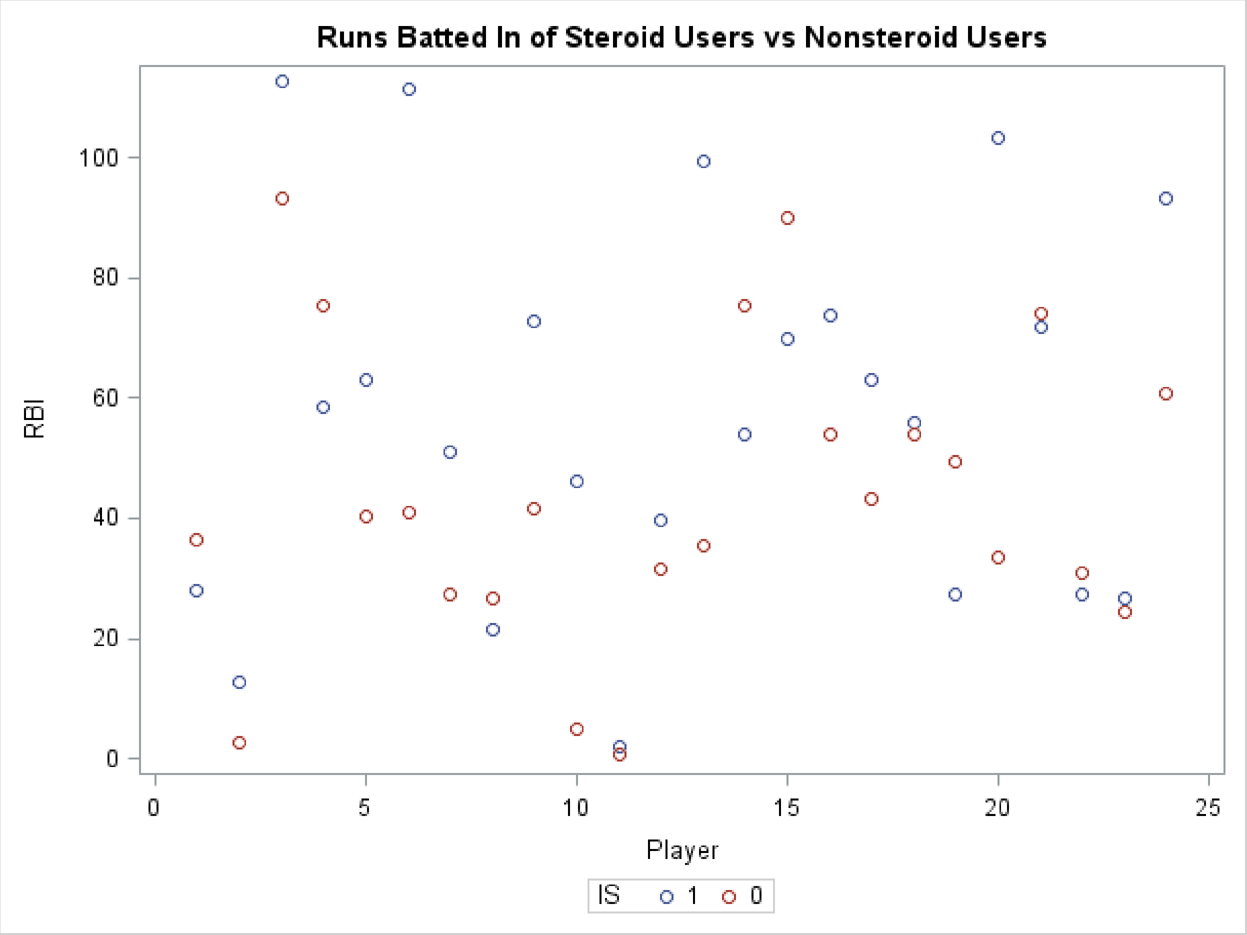

Figure 2.2: RBIs of Steroid Users vs Nonsteroid Users

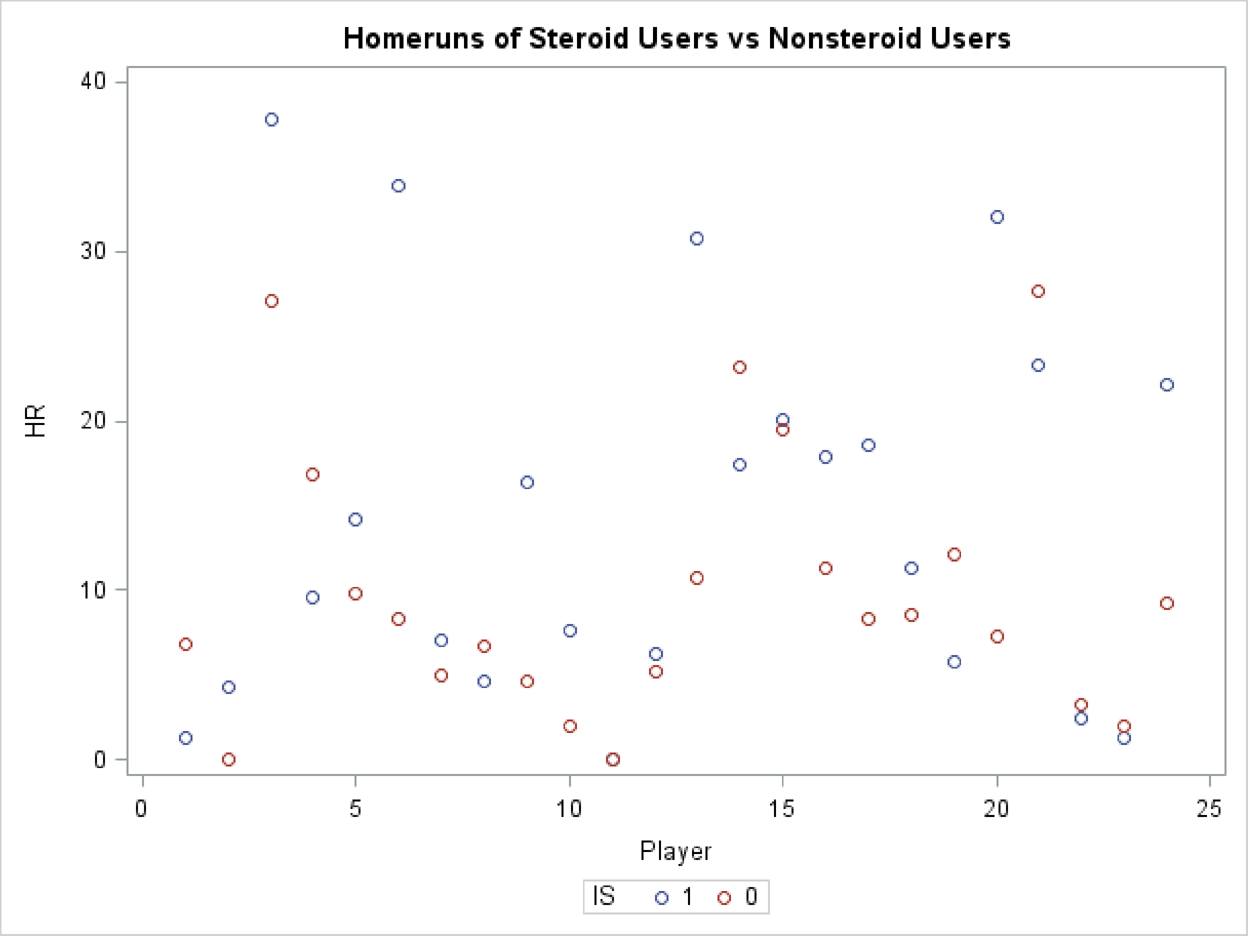

Figure 2.3: Homeruns of Steroid Users vs Nonsteroid Users

also known as the Allocation Rule (Everitt, 2010).

Linear discriminant analysis allows one to find the line that best separates two groups of data. Another way to look at it is by understanding that discriminant function analysis is the same as MANOVA reversed. MANOVA stands for multivariate analysis of variance. Discriminant analysis uses the explanatory variables as the predictors while the response variables are the groups. MANOVA is just the opposite. The groups are the explanatory variables while the predictors are the response variables.

In addition, the assumptions of MANOVA also apply to discriminant analysis. The first assumption is that the sample sizes of the groups being analyzed are not necessarily required to be equal. Next, the data are assumed to follow a multivariate normal distribution. However, the analysis is not inaccurate if the data do not follow this distribution, as long as non-normality is caused by skewness rather than outliers. The variances and covariances are assumed to be homogeneous among the groups being studied. Homogeneous variances and covariances can be achieved by transforming the variables. Extreme outliers should be eliminated or transformed. Their presence affects the mean and causes increased variability. The last assumption is that there is low multicollinearity of the potential explanatory variables. If a potential explanatory variable is highly correlated to another or even a function of another, then the matrix will not have a unique discriminant solution (Poulsen, 2008).

Discriminant Analysis serves two primary purposes. First, it is a forecasting function, meaning that it is used to predict the correct group for individuals not included in the equation-determining sample. The second purpose it serves is to act as an evaluative function. As an evaluative function, it examines how well the function fits the data from which it was created (Marascuilo, 1983).

Logistic regression is sometimes used in place of discriminant analysis because it can answer the same types of problems. Most people actually prefer logistic regression because it is more flexible. It can handle continuous and categorical variables while discriminant analysis deals only with continuous variables (Poulsen, 2008). Logistic regression will be used in combination with linear discriminant analysis for this thesis.

By looking at Figures 2.1, 2.2, and 2.3, we can see that there are differences between the two groups, but a linear discriminant function will show us which variables are important in determining the distinction between players who have been suspended for PED use and players who have never been suspended for PEDs. The function developed can then be used to distinguish between the individual players in the suspected group of players.

The process of finding the best discriminant function for this particular data set starts with logistic regression. However, since the ultimate goal is to obtain a linear discriminant function, we start by removing any variables that are collinear with other variables. Some of the variables shown in Table 2.1 are functions of other variables. The batting average of a player, for instance, is calculated by dividing his number of hits (H) by his number of at bats (AB). Therefore, in this case, hits and at bats are not used in the logistic regression model.

After removing all collinear variables, there are two groups of non-collinear variables that can possibly be used in the linear discriminant function. Two groups exist because OPS and OPS adjusted to the player's ballpark can remain in the discriminant function if slugging percentage and on-base percentage are removed, and vice versa. As mentioned before, some variables are functions of other variables. OPS is the sum of On-Base Percentage and Slugging Percentage. The formula for calculating the OPS adjusted to the player's ballpark is as follows:

$$OPSplus = 100 * [\frac{OBP}{LeagueOBP} + \frac{SLG}{LeagueSLG} - 1] \tag{2.2}$$

Since both OPS and OPS adjusted to the player's ballpark are calculated by formulas that include On-Base Percentage and Slugging Percentage, OPS and OPS adjusted to the player's ballpark cannot be contained in the same discriminant function as On-Base Percentage and Slugging Percentage. Therefore, group one includes Games, Plate Appearances, Runs, Runs

Batted In, Stolen Bases, Caught Stealing, Strikeouts, Batting Average, On-Base Percentage, Slugging Percentage, Grounded Double Plays, Sacrifice Hits, and Intentional Walks. The second group consists of Games, Plate Appearances, Runs, Runs Batted In, Stolen Bases, Caught Stealing, Strikeouts, Walks, Batting Average, OPS, OPS adjusted to the player's ballpark, Total Bases, Grounded Double Plays, Hit By Pitch, Sacrifice Flies, and Intentional Walks. Each of these groups of variables will be considered when creating discriminant functions.

Although including all variables of each group would probably produce two very accurate formulas, the formulas would be overparameterized. Therefore, we use both forward selection and backward selection on each group to properly select four models. These four logistic regression models are listed below.

$$IS = 3.862 - 9.295SLG \tag{2.3}$$

Model 2.3 is the result of forward selection on group one. It has an AIC of 65.71 and an SC of 69.45. The AIC and SC are criteria for selecting the best model from a finite set of models. AIC, or Akaike Information Criterion, is calculated using the formula $2k - 2ln(L)$ where k is the number of parameters contained within the model and L is the maximum value of the likelihood function. SC, or Schwarz Criterion, is also known as the Bayesian Information Criterion, or BIC. It is calculated using the formula BIC$= -2ln(L) + k(ln(n))$, where n is the sample size. A model with the smallest AIC and/or BIC values is typically the best choice (methodology.psu.edu). Slugging Percentage is significant at the $\alpha = 0.05$ level for this model.

$$IS = 9.886 + 0.054BB - 15.88OPS \tag{2.4}$$

Model 2.4 is the result of backward selection on the first group of variables. It has an AIC of 62.91 and an SC of 68.52, therefore, based on these criteria, Model 2.4 is the better choice if only selecting the best model from models 2.3 and 2.4. Each of the parameters are significant at the $\alpha = 0.05$ level for this model as well.

$$\begin{aligned} IS &= 33.84 + 0.1897G - 0.071PA + 0.54R - 0.549RBI \\ &+ 0.072SO - 127.3OBP + 0.672GDP - 1.0497SH + 1.767IBB \end{aligned} \tag{2.5}$$

Model 2.5 results from forward selection on the second group of variables. It has an AIC of 50.61 and an SC of 69.33. If using only the AIC to determine the best model out of these first three models, Model 2.5 would be the best model. However, if using SC, Model 2.4 would be the best model to use. Each of the parameters are significant at the $\alpha = 0.05$ level for this model.

$$IS = 11.20 + 0.078R - 0.0895RBI - 45.368BA + 0.371IBB \tag{2.6}$$

Model 2.6 is the result of backward selection on group two. It has an AIC of 60.74 and an SC of 70.13 so Model 2.6 would not be the best model based on either of these criteria. Like the first three models, each of the parameters are significant at the $\alpha = 0.05$ level. Based on its AIC, we will choose Model 2.5 as the best logistic regression model.

Now that we have four logistic regression models, we can test their accuracy at classifying steroid and non-steroid users by performing linear discriminant analysis using each model. First we will see how well Model 2.5 classifies the steroid users and the non-steroid users, as shown in Table 2.2. Since we are estimating IS, a variable that can only have one of two values, 0 or 1, an estimated IS that is greater than 0.5 will indicate a player who is classified as a steroid user. An estimated IS that is close to zero, i.e. less than 0.5, will indicate that the player is classified as a non-steroid user. Model 2.5 accurately classifies thirty-six out of forty-eight players, with an error rate of 25%.

By standardizing the data and then performing linear discriminant analysis, we now have

Table 2.2: Model 2.5: Assigned Class vs True Class

Player	Estimated IS	Assigned	True Class	Player	Estimated IS	Assigned	True Class
1	0.913	1	1	25	0.146	0	0
2	0.474	0	1	26	0.489	0	0
3	0.960	1	1	27	0.504	1	0
4	0.712	1	1	28	0.508	1	0
5	0.421	0	1	29	0.314	0	0
6	0.633	1	1	30	0.104	0	0
7	0.856	1	1	31	0.362	0	0
8	0.770	1	1	32	-0.056	0	0
9	0.642	1	1	33	0.450	0	0
10	0.484	0	1	34	0.273	0	0
11	0.478	0	1	35	-0.166	0	0
12	0.379	0	1	36	0.641	1	0
13	1.006	1	1	37	0.333	0	0
14	0.454	0	1	38	0.117	0	0
15	0.648	1	1	39	0.558	1	0
16	0.854	1	1	40	0.345	0	0
17	0.588	1	1	41	0.207	0	0
18	0.474	0	1	42	-0.265	0	0
19	1.022	1	1	43	0.448	0	0
20	1.147	1	1	44	-0.253	0	0
21	0.818	1	1	45	0.496	0	0
22	0.638	1	1	46	0.059	0	0
23	1.179	1	1	47	0.464	0	0
24	0.694	1	1	48	0.680	1	0

four new functions that relate to each of the four logistic regression models, respectively. These discriminant functions are shown below.

$$IS = 1.0397SLG \tag{2.7}$$

$$IS = -1.1916BB + 1.691OPS \tag{2.8}$$

$$\begin{aligned} IS &= -2.667G + 4.509PA - 3.901R + 3.751RBI \\ &- 0.527SO + 0.86OBP - 0.899GDP + 0.705SH - 1.086IBB \end{aligned} \tag{2.9}$$

$$IS = -1.669R + 1.865RBI + 1.007BA - 0.786IBB \tag{2.10}$$

All four functions are cross-validated using leave-one-out cross-validation in order to ensure stability. Leave-one-out cross-validation is a technique used for determining a function's true error rate. During leave-one-out cross-validation, a model is repeatedly refit leaving out one observation at a time and is then used for deriving a prediction of that left-out observation. Without cross-validation, the model will in some way overfit the data. For example, before cross-validation, Function 2.9 accurately classifies seventeen steroid users and accurately classifies nineteen non-steroid users. Table 2.3 shows the classification ability of Function 2.9 before cross-validation, and Table 2.4 shows the classification ability of Function 2.9 after cross-validation.

Table 2.3 shows that, using Function 2.9, nineteen non-steroid users classify as non-steroid users, but five non-steroid users are classified as steroid users. It also shows that seventeen steroid users are classified as steroid users, but seven steroid users are classified as non-steroid

Table 2.3: Function 2.9 Classification Prior to Cross-Validation

	0	1
0	19	5
1	7	17

Table 2.4: Function 2.9 Classification After Cross-Validation

	0	1
0	15	9
1	8	16

users. This gives an error rate of (7+5)/48=25% for Function 2.9 before cross-validation. Table 2.4 shows that, after cross-validation, Function 2.9 inaccurately classifies nine non-steroid users as steroid users and eight steroid users as non-steroid users, giving a new, more precise error rate of 35.4%.

Table 2.5 shows that, using Function 2.10, thirteen non-steroid users classify as non-steroid users, but eleven non-steroid users are classified as steroid users. It also shows that eighteen steroid users are classified as steroid users, but six steroid users are classified as non-steroid users. This gives an error rate of (6+11)/48=35.4% for Function 2.10 before cross-validation. Table 2.6 shows that, after cross-validation, Function 2.10 inaccurately classifies twelve non-steroid users as steroid users and seven steroid users as non-steroid users, giving a more precise error rate of 39.6%.

Now that we have four discriminant functions, it is important to test how accurate each function is at classifying each player into his proper group. One way to examine a function's

Table 2.5: Function 2.10 Classification Prior to Cross-Validation

	0	1
0	13	11
1	6	18

Table 2.6: Function 2.10 Classification After Cross-Validation

	0	1
0	12	12
1	7	17

accuracy is to look at the plot of its discriminant scales. A plot such as this shows how well the function separates the two groups of data.

Figures 2.4, 2.5, 2.6, and 2.7 show that none of the four discriminant functions is particularly good at separating steroid users from nonsteroid users. If one had to choose the best function based solely on these plots, the best function would be Function 2.9 with Function 2.10 following close behind. These two functions produce the least amount of overlap when separating the two groups of data.

Table 2.7 shows the true class and the assigned class of each player based on Function 2.9's implementation. Again, Class 1 represents PED users while Class 0 represents players never suspended or suspected of PEDs. It also gives the posterior probabilities obtained for each of the players. For example, the posterior probability that player one is of Class 0 is 0.067, and the posterior probability that he is of Class 1 is 0.933. This was an accurate classification because he was, in fact, suspended from Major League Baseball in April of 2005.

Looking at Table 2.7, it is apparent that Function 2.9 does not accurately classify each of the first forty-eight players. To be exact, Function 2.9 accurately classifies thirty-one out of forty-eight players. This gives an error rate of 35.4%. Table 2.8 shows that Function 2.10 also does not correctly classify all of the players from the groups of steroid users and nonsteroid users. Function 2.10 accurately classifies twenty-nine out of the forty-eight players with an error rate of 39.6%, as discussed before. Although the error rate of the latter function is higher, it accurately classifies seventeen out of the twenty-four suspended players while Function 2.9 correctly classifies sixteen out of the twenty-four suspended players. Since the

Table 2.7: Function 2.9: Assigned Class vs True Class

Player	P(Class 0)	P(Class 1)	True Class	Player	P(Class 0)	P(Class 1)	True Class
1	0.067	0.933	1	25	0.876	0.124	0
2	0.728	0.272	1	26	0.359	0.641	0
3	0.046	0.954	1	27	0.338	0.662	0
4	0.216	0.784	1	28	0.229	0.771	0
5	0.784	0.216	1	29	0.731	0.269	0
6	0.695	0.305	1	30	0.927	0.073	0
7	0.140	0.859	1	31	0.696	0.304	0
8	0.174	0.826	1	32	0.978	0.022	0
9	0.431	0.569	1	33	0.524	0.476	0
10	0.704	0.296	1	34	0.765	0.235	0
11	0.702	0.298	1	35	0.998	0.002	0
12	0.825	0.175	1	36	0.202	0.798	0
13	0.032	0.968	1	37	0.725	0.275	0
14	0.691	0.309	1	38	0.920	0.08	0
15	0.437	0.563	1	39	0.272	0.728	0
16	0.094	0.906	1	40	0.557	0.446	0
17	0.394	0.606	1	41	0.851	0.149	0
18	0.591	0.409	1	42	0.9998	0.0002	0
19	0.027	0.973	1	43	0.412	0.588	0
20	0.009	0.991	1	44	0.997	0.003	0
21	0.123	0.877	1	45	0.457	0.543	0
22	0.493	0.507	1	46	0.942	0.058	0
23	0.006	0.994	1	47	0.432	0.568	0
24	0.352	0.648	1	48	0.052	0.948	0

Table 2.8: Function 2.10: Assigned Class vs True Class

Player	P(Class 0)	P(Class 1)	True Class	Player	P(Class 0)	P(Class 1)	True Class
1	0.407	0.593	1	25	0.354	0.646	0
2	0.292	0.708	1	26	0.861	0.139	0
3	0.185	0.815	1	27	0.410	0.590	0
4	0.446	0.554	1	28	0.372	0.628	0
5	0.669	0.330	1	29	0.359	0.641	0
6	0.433	0.567	1	30	0.448	0.552	0
7	0.657	0.343	1	31	0.519	0.481	0
8	0.542	0.458	1	32	0.862	0.138	0
9	0.251	0.749	1	33	0.559	0.441	0
10	0.729	0.271	1	34	0.959	0.041	0
11	0.720	0.280	1	35	0.987	0.013	0
12	0.274	0.726	1	36	0.312	0.688	0
13	0.475	0.525	1	37	0.415	0.585	0
14	0.144	0.856	1	38	0.616	0.384	0
15	0.575	0.425	1	39	0.481	0.519	0
16	0.153	0.847	1	40	0.370	0.630	0
17	0.343	0.657	1	41	0.470	0.530	0
18	0.318	0.682	1	42	0.993	0.007	0
19	0.272	0.728	1	43	0.242	0.758	0
20	0.084	0.916	1	44	0.810	0.189	0
21	0.190	0.810	1	45	0.613	0.387	0
22	0.869	0.130	1	46	0.842	0.158	0
23	0.207	0.793	1	47	0.526	0.474	0
24	0.246	0.754	1	48	0.449	0.551	0

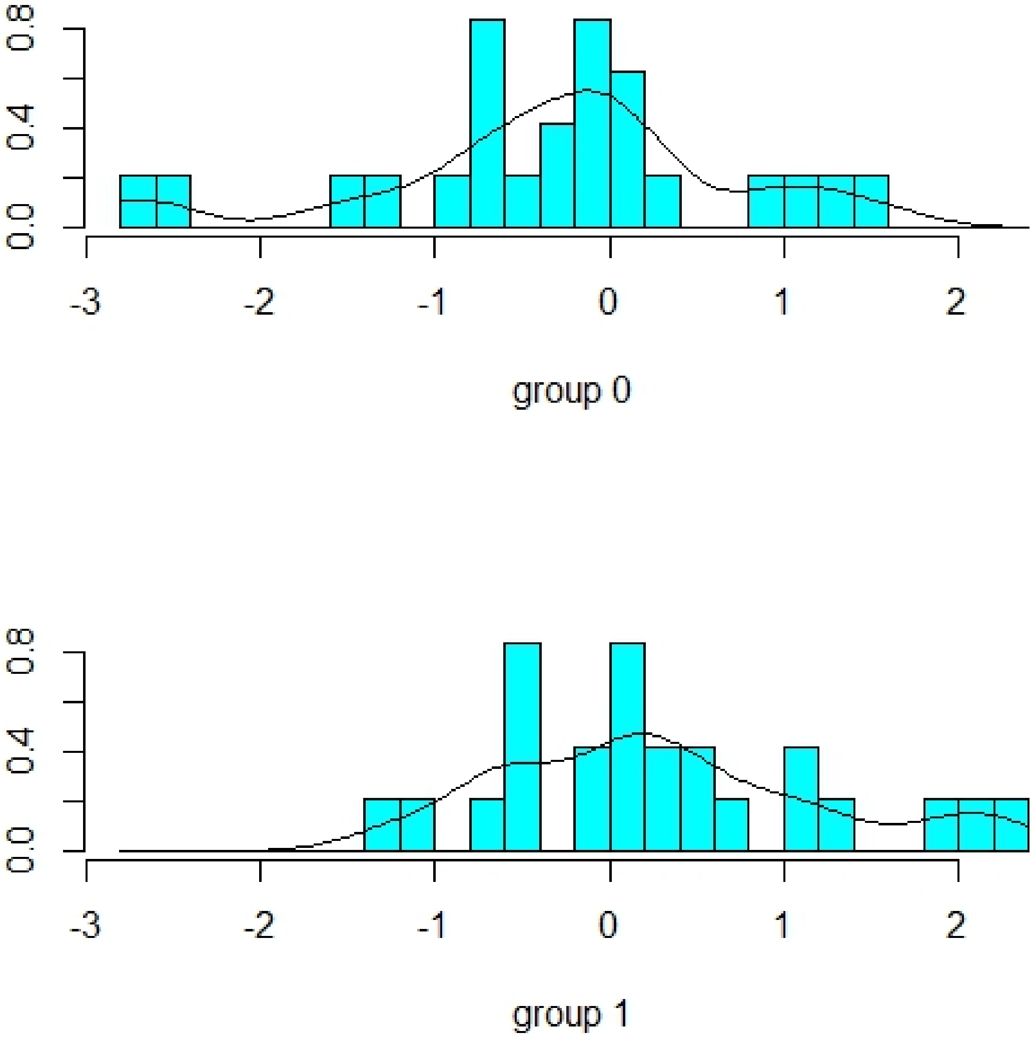

Figure 2.4: Function 2.7's Discriminant Scales of Steroid Users vs Nonsteroid Users

suspended players are definitely steroid users, Function 2.10 might be a more reliable function to use for classifying the suspected players. Also, Function 2.9 contains nine variables as opposed to Function 2.10's four variables so it may be slightly overparameterized. Therefore,

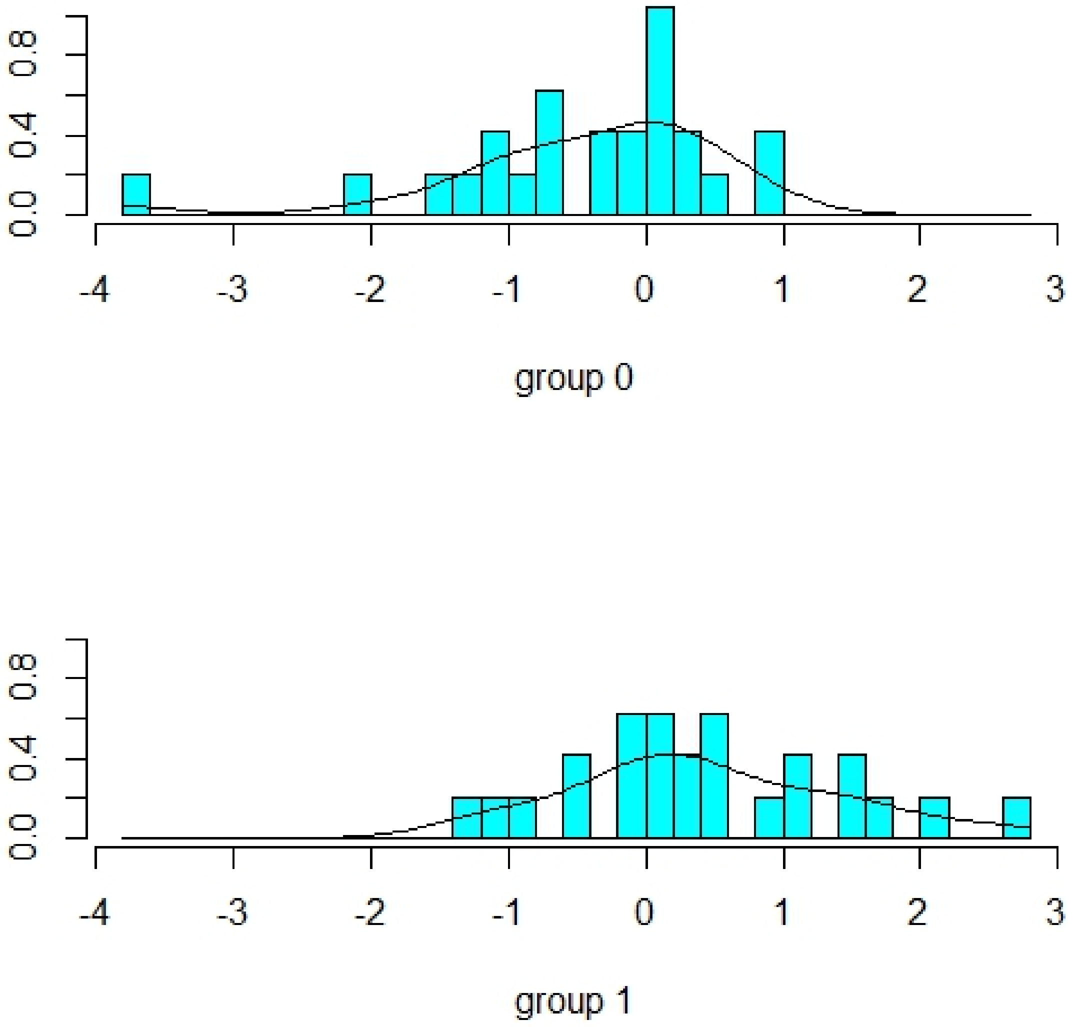

Figure 2.5: Function 2.8's Discriminant Scales of Steroid Users vs Nonsteroid Users

Function 2.10 is chosen as the best discriminant function for determining steroid users.

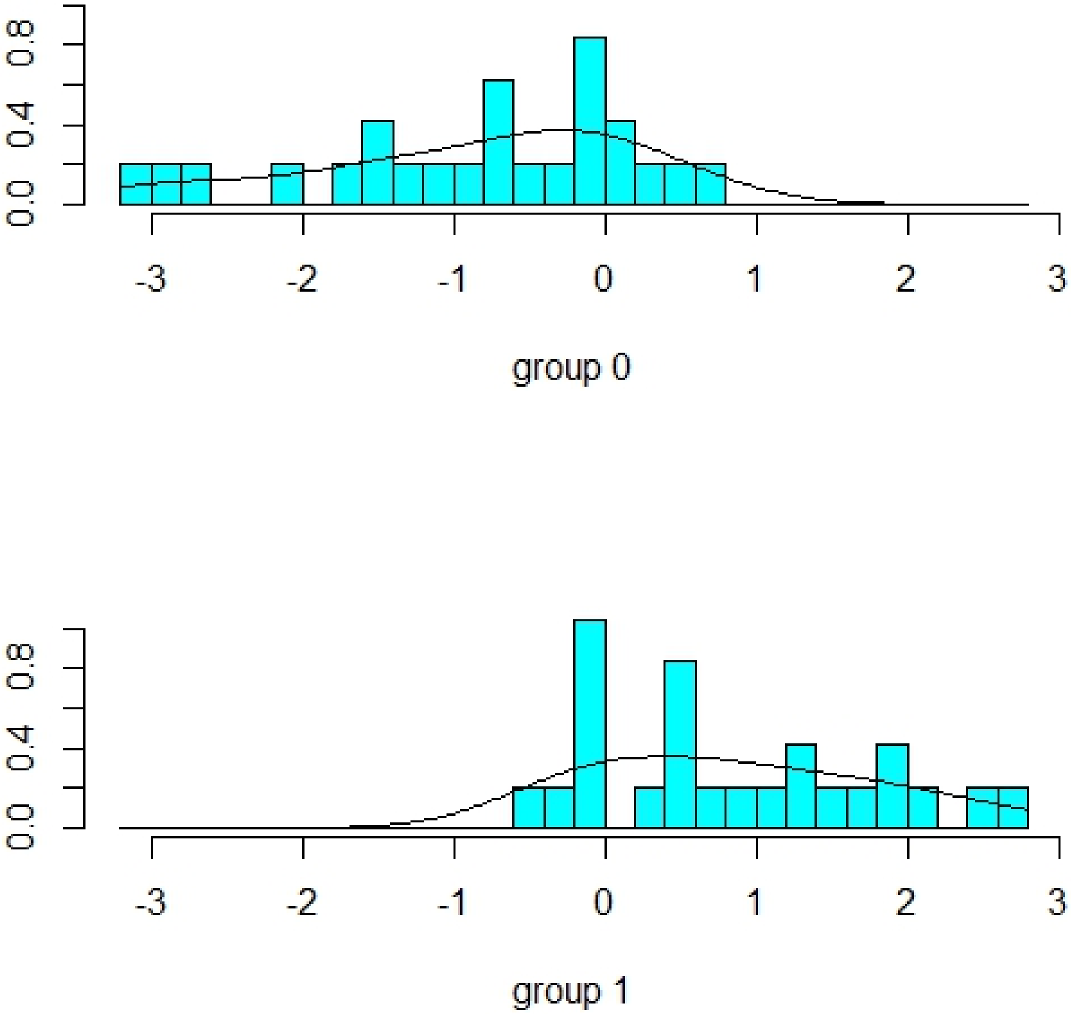

Figure 2.6: Function 2.9's Discriminant Scales of Steroid Users vs Nonsteroid Users

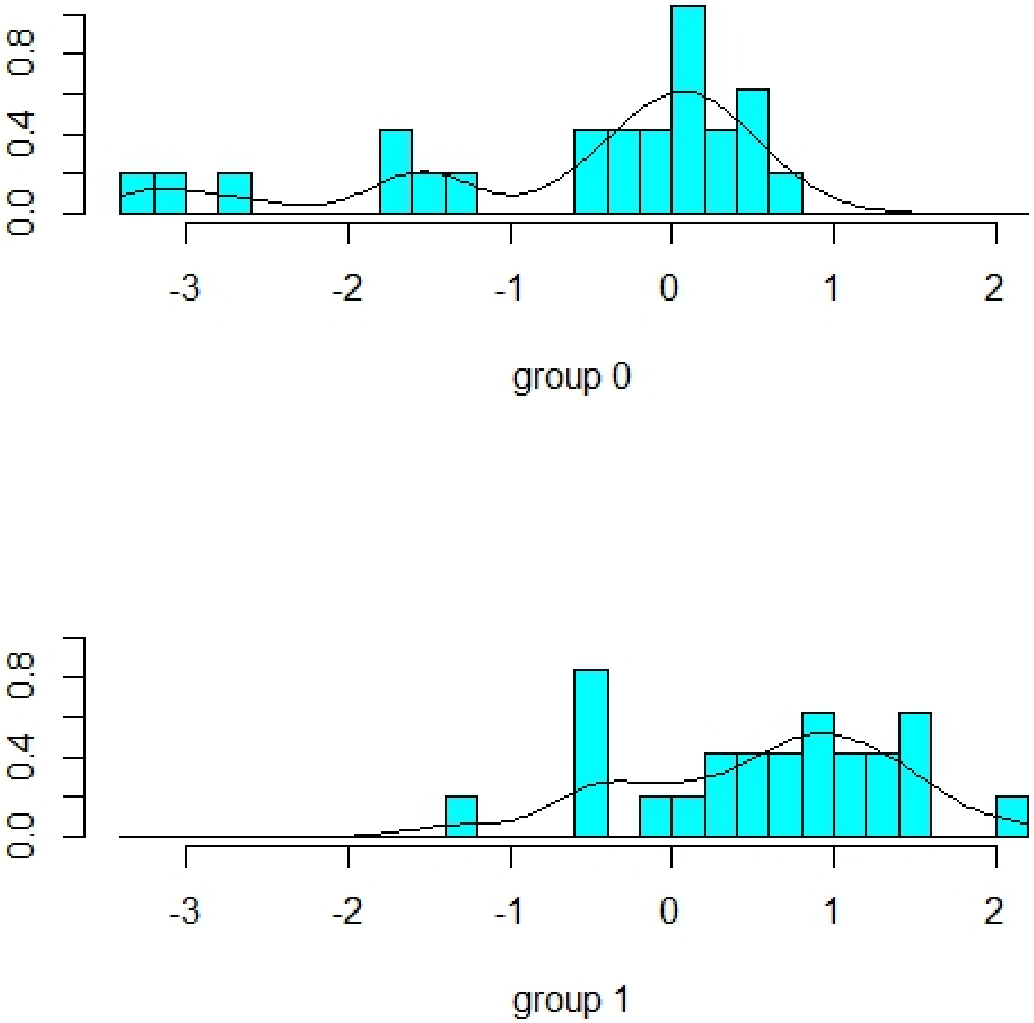

Figure 2.7: Function 2.10's Discriminant Scales of Steroid Users vs Nonsteroid Users

Chapter 3

Results

3.1 Results of Suspected Players

According to our best linear discriminant function (Function 2.10),sixteen out of the thirty-four players who were suspected of steroids match the criteria of steroid users. The results are shown in Table 3.1. For the protection of the players' privacy, their names are not mentioned. The posterior probabilities for the assigned classes are almost always large with the exception of just a few players. For instance, player 1 has a probability of 0.826 of belonging to Class 0 and a 0.174 probability of belonging to Class 1. It appears that the discriminant function is quite certain that this player is not a steroid user. The closest probabilities belong to player 25 who has a 0.514 probability of belonging to Class 0 and a 0.486 probability of belonging to Class 1.

According to our best logistic regression model (Model 2.5), sixteen out of the thirty-four players who were suspected of steroids match the criteria of steroid users. The results are shown in Table 3.2.

Table 3.1: Function 2.10: Assigned Classes of Suspected Players

Player	P(Class 0)	P(Class 1)	Assigned Class
1	0.826	0.174	0
2	0.799	0.200	0
3	0.965	0.035	0
4	0.283	0.717	1
5	0.890	0.110	0
6	0.433	0.567	1
7	0.637	0.363	0
8	0.687	0.313	0
9	0.844	0.156	0
10	0.380	0.620	1
11	0.706	0.294	0
12	0.459	0.541	1
13	0.041	0.959	1
14	0.781	0.219	0
15	0.221	0.779	1
16	0.684	0.316	0
17	0.719	0.281	0
18	0.620	0.380	0
19	0.276	0.724	1
20	0.558	0.442	0
21	0.217	0.783	1
22	0.643	0.357	0
23	0.560	0.440	0
24	0.423	0.577	1
25	0.486	0.514	1
26	0.168	0.832	1
27	0.036	0.964	1
28	0.506	0.494	0
29	0.840	0.160	0
30	0.174	0.826	1
31	0.219	0.781	1
32	0.210	0.790	1
33	0.329	0.671	1
34	0.535	0.465	0

Table 3.2: Model 2.5: Assigned Classes of Suspected Players

Player	Estimated IS	Assigned Class
1	0.015	0
2	- 0.337	0
3	-1.220	0
4	0.646	1
5	0.649	1
6	0.718	1
7	0.404	0
8	0.331	0
9	0.048	0
10	-0.118	0
11	0.088	0
12	0.719	1
13	1.166	1
14	0.246	0
15	0.591	1
16	0.127	0
17	0.377	0
18	0.767	1
19	0.995	1
20	0.607	1
21	0.797	1
22	0.191	0
23	0.346	0
24	0.766	1
25	0.439	0
26	0.447	0
27	1.136	1
28	0.633	1
29	0.561	1
30	0.427	0
31	0.515	1
32	0.599	1
33	0.450	0
34	0.572	1

3.2 Thoughts and Concerns

While the discriminant function appears to report accurate results approximately 60% of the time, there are still some concerns that need to be addressed. First, if the players who have never been suspended or suspected of using performance-enhancing drugs have truly never used them during the duration of their entire careers, then this is an accurate function. However, we cannot be completely certain that this is true, and there is no way we will really know unless the players one day confess to having used the drugs.

As briefly mentioned before, Tony Bosch admitted to personally injecting Alex Rodriguez with performance-enhancing substances in January of 2014. He also stated that Alex Rodriguez would put testosterone troches in his mouth before the game started and “by the time they get back into the locker room and there was any possibility of testing, they would test clean." (Barry, 2014) If MLB players have actually mastered the art of taking these drugs and some of the players in the group designated as non-steroid users have used steroids at some point in their career, their batting statistics are not entirely clean. This would then produce inaccurate results.

Assuming, for now, that the players in that group have all-natural batting statistics, we still have one other concern at hand. We cannot control for talent. Some MLB players are naturally more talented than others. This is true for all players in every sport and of every age, from little league baseball to the National Football League to the National Basketball Association. Due to the fact that talent cannot be controlled for when analyzing the performance of athletes, the discriminant function and logistic regression model cannot necessarily be used to provide reason for suspending a player from Major League Baseball. However, maybe they could be used in conjunction with drug testing. These formulas could help bring certain individuals to the attention of MLB, therefore allowing it to monitor every player’s performance. If he is classified as a steroid user based on Function 2.10 or Model 2.5, perhaps he should be drug tested more frequently or further investigated.

Chapter 4

Conclusion

Although drug testing has been present in Major League Baseball since 2003, MLB players continue to violate rules against steroids and other performance-enhancing drugs year after year. The purpose of this study was to find an alternative and possibly more effective method for determining and ultimately preventing steroid use in the MLB. This method is based on batting statistics of only position players.

After some initial exploratory data analysis, it was decided that linear discriminant analysis and logistic regression would be the best approaches to this issue since we want to distinguish the group of steroid users from the group of players who have never been suspended for steroids. The analysis resulted in four formulas using each method. One formula was selected from each method based on its ability to classify the players. The formulas were then applied to our group of suspected users of steroids. At this point, both formulas predicted that sixteen out of thiry-four players from this group are steroid users.

According to Function 2.10, the best variables for classifying a player as a steroid user are Runs, Runs Batted In, Batting Average, and Intentional Walks. As mentioned before, Slugging Percentage, Runs Batted In, and Homeruns were expected to be decent determining factors of a steroid user. It is surprising that Slugging Percentage was not included in either formula, considering it is the most discussed statistic when speaking of performance-enhancing drugs. Runs and Batting Average are not surprising since good athletes tend to perform better in these areas. Intentional Walks was not a variable that was expected to be included in the formula, however, it is not entirely surprising that it is included since batters are intentionally walked for strategic reasons. A pitcher will usually walk a batter intentionally if that batter is expected to perform well due to a high batting average, slugging percentage, amount of homeruns, etc.

If the formulas are accurate and sixteen out of thirty-four MLB players suspected of steroid use actually used steroids or some type of performance-enhancing drug at some point in their career, what does this say about the rest of the MLB players? Again, the formulas should not be used for suspending players or even accusing players of steroid use, but they could potentially be used by Major League Baseball for determining which players the organization needs to watch more closely. The road to using statistics for resolving MLB's steroid issue is still a long one, but perhaps these new formulas are a good start to a new beginning for Major League Baseball.

Bibliography

[1] Barry, J. “Anthony Bosch: A-Rod knew he was taking banned substances." *Miami Herald*. January 13, 2014. March 2014. <http://www.miamiherald.com/2014/01/12/3867470/anthony-bosch-a-rod-knew-they.html>.

[2] “A Complete Application: Analysis of the Fisher Iris Dataset." *Idiap Research Institute.* March 2014. <www.idiap.ch/software/bob/docs/releases/last/sphinx/html/IrisExample.html>.

[3] Everitt, B., and A. Skrondal. *The Cambridge Dictionary of Statistics.* 4th ed. New York: Cambridge University Press, 2010.

[4] Poulsen, J. and A. French. “Discriminant Function Analysis." *San Francisco State University.* June 3, 2008. March 2014. <http://userwww.sfsu.edu/ efc/classes/biol710/discrim/discrim.pdf>.

[5] Kendrick, S. “Baseball Players Accused Of Using Performance-Enhancing Drugs." *About.* March 2014. <http://baseball.about.com/od/majorleagueplayers/a/drugplayers.htm>.

[6] “List of Players Linked to Steroids and/or Human Growth Hormone." *Baseball’s Steroid Era.* August 9, 2006, March 2014. <http://thesteroidera.blogspot.com/2006/08/list-of-steroid-hgh-users-in-baseball.html>.

[7] Marascuilo, L., and J. Levin. *Multivariate Statistics in the Social Sciences.* California: Brooks/Cole Publishing Company, 1983.

[8] Marchi, M., and J. Albert, *Analyzing Baseball Data with R.* Florida: Taylor & Francis Group, LLC, 2014.

[9] "MLB List of Users of PEDs & Suspected Users of PEDs." *Fannation.* November 2013. <http://www.fannation.com/blogs/post/265355-mlb-list-of-users-of-peds-amp-suspected-users-of-peds>.

[10] Rymer, Z. "Full Timeline of MLB's Failed Attempts to Rid the Game of PEDs." *Bleacher Report.* June 10, 2013. March 2014. <http://bleacherreport.com/articles/1667581-full-timeline-of-mlbs-failed-attempts-to-rid-the-game-of-peds>.

[11] "AIC vs. BIC."*The Methodology Center.* 2007. April 2014. <http://methodology.psu.edu/eresources/ask/sp07t>.