A STATISTICAL APPROACH TO PREDICT FUTURE MEMBERS OF THE BASEBALL HALL OF FAME

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF MATHEMATICS

IN THE GRADUATE SCHOOL OF THE

TEXAS WOMAN'S UNIVERSITY

DEAPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

COLLEGE OF ARTS AND SCIENCES

BY

KELSEY ZEMLER, B.S.

DENTON, TEXAS

DECEMBER 2013

ACKNOWLEDGEMENTS

I want to sincerely thank the entire mathematics and computer science faculty for teaching me throughout my master's degree. To my advisory committee, thank you for guiding me throughout my classes as well as my thesis research.

I would like to acknowledge my family and friends for their love and support as I continued my education. To my loving husband, Keene, who helped me through the long process of obtaining my degree and completing my thesis. And to my baby girl, Leah, who will, one day, appreciate math and baseball as much as I do.

ABSTRACT

KELSEY ZEMLER

A STATISTCAL APPROACH TO PREDICT FUTURE MEMBERS OF THE BASEBALL HALL OF FAME

DECEMBER 2013

The purpose of this study was to construct an accurate statistical model for the members of the baseball hall of fame and use this model to predict future hall of fame members, using a frequentist approach. Using logistic regression, accurate models were determined for each position that can be used to predict if a certain player will make it into the Hall of Fame. Baseball-Reference.com was the major source of data. Once the analysis was complete, nine different models were chosen to determine the probability that a certain player at a given position would make it into the Hall of Fame based on their baseball statistics. Positions were also analyzed by time periods and models were found for each position in each time frame, if one existed. In general, time period models for the various positions were inconclusive, however a model for each position was found overall.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

Figure PAGE

CHAPTER I

REVIEW OF LITERATURE

To some, baseball has English roots known as Rounders, which in turn became known as town ball (Durant, p. 2). For many, it is pure American. Investigations were done in the early 1900's and Mills reported that baseball did originate in America and the first known time of playing is in 1839 by Doubleday. Which story is correct? The most accepted form was that it came to America based on Rounders, and has since developed into America's national pastime. The Knicks was the first organized ball team and in 1846, the game of modern baseball erupted. In fact many aspects are visible today: 9 man team, and flat bases 90 feet apart in the shape of a diamond (Durant, p.6).

Major League Baseball is comprised of two leagues: American and National. The National League came into existence in 1876 as a reform effort to get rid of gambling, among other things, in professional baseball. Hulbert became known as the creator of the National League. In 1900, the American League came into existence thanks to Byron Johnson (Durant, p.45).

After over 100 years, baseball has become larger than ever. The National Baseball Hall of Fame and Museum is located in Cooperstown, and houses numerous photos, letters, and memorabilia from baseball's past (Durant, p. 45-6). The Hall of Fame exists

within the Museum and is home to 237 players, in addition to many umpires, managers, and executives.

According to baseballhall.org, there are many requirements that a candidate must fulfill in order to be inducted into the hall of fame. In order to be inducted, a player must have played in the majors within the frame of 20 years before and 5 years prior to election. A player must have played in 10 championship seasons and some of them have to be in the time frame described above. If an active player dies, they are eligible for the hall 6 months after death or after the end of the 5 years period. If a player meets these requirements, they must receive a 75% vote by the election committee (BBWAA).

There are currently 2 ways that a player can be inducted into the Hall of Fame: the Baseball Writers Association of America (which was described above), and the Veterans Committee. The Veterans committee is comprised of all members of the Hall of Fame and still requires a 75% vote. The Veterans committee also elects the players from the Negro League as of 1977. Prior to this, the Negro League nominated players. Starting in 1939, players could also be voted in by the Old Timers Committee. Executives and sports writers voted for the players into entry, but this practice later stopped when the Veterans Committee was created. In 2005, a special committee for African American Baseball's election process was approved by the Hall of Fame and was able to elect members for 2006. (Hall of Fame Voting Procedures)

As always, if a player has been placed on the ineligible list, they cannot be inducted into the Hall of Fame. After looking at the ineligible list (Baranger, 2007), the players and umpires listed were involved in gambling, intentionally throwing games, or drug related offenses. Although many players have been placed on the ineligible list, many have also been reinstated.

Statistics in baseball has been around since the beginning, however sabermetrics, as it is called, is a new trend. Made famous by Bill James in the 1980's, sabermetrics is the statistical study of baseball. Sabermetrics uses statistical analysis to evaluate players and answer theories/questions (Birnbaum).

Although statistics is nothing new to baseball, the developing ideas of sabermetricians are. In Bill James' book "*The Bill James Baseball Abstract 1988,"* he wrote about a few of his discoveries, some of which are listed below: "Talent in baseball is not normally distributed. It is a pyramid…Ballplayers, as a group, reach their peak value much earlier and decline much more rapidly than people believe…The chance of getting a good player with a high draft pick is substantial enough that it is clearly a disastrous strategy to give up a first round draft pick to sign a mediocre free agent…" (James, 1988).

According to Ted Johnson, there are 5 major time frames in which baseball changed drastically. Johnson looked at the composition of the baseball to create the eras. The first time period he created was Pre-1872. The baseball at this time was made of a

rubber and string core, wrapped in horse hide. Weight fluctuated and was inconsistent in size. According to Johnson, it "played like it was stuffed with feathers…"

The next time frame started in 1872; Standards. The make of the ball was the same, however the ball was standardized: 9.25 inches in circumference, weighed 5 – 5.25 ounces. Despite the standardization of the dimensions of the baseball, the ball still played as it did pre-1872.

The third time frame was characterized by a cork center in 1910. Because of the center, the ball went further. The "dead ball" era, 1872-1910, was over and the "live ball" era emerged.

The fourth time frame changed the ball dramatically. The cork center was still in place; however rubber was wrapped around it. The seams of the ball were also raised for easier release. This time period definitely helped pitchers: the raised seam enabled pitchers to increase pitching repertoires.

The final time frame created started in 1974. This time frame is characterized by the ball no longer being wrapped in horse hide. Due to a lack of horse hide, cowhide was now implemented. This time period is still in existence today (Johnson T.).

When running logistic regressions, cross-validating the model created is usually an essential part of the analysis. According to Picard and Cook, there are both pros and cons to cross validation, as well as circumstances in which it does not apply. Data

splitting, in which the data is split into two groups (one for model building, the other for model testing) has been met with controversies. “The loss of information incurred in model development…” was a concern. (Richard Picard, 1984) It is reasonable to split large data sets and cross validate. However for small data sets, this could lead to an unreliable model.

Data splitting for cross validation is widely practiced, and is sometimes overdone. It can lead to many problems, which was addressed above. According to Mosteller and Tukey, “Testing the procedure on the data that gave it birth is almost certain to overestimate performance…” (Richard Picard, 1984) According to this statement, cross-validation is *not* a good method for testing model fit for population data.

There have been many studies that are all in search to obtain an accurate model for entry into the baseball Hall of Fame. Many students in academia have also studied hall of fame entry. In “Using Binary Logisitic Regression to Predict Baseball Hall of Fame Admittance,” Adam Grajkowske used logistic regression to evaluate players entry into the hall of fame based on being a hitter or being a pitcher. Grajkowske selected a small number of variables for hitters and pitchers. Using minitab, logistic regressions were run for both. Hosmer-Lemeshow tests were conducted for goodness of fit and two models were obtained. Results were given and models were compiled. (Grajkowske, 2008)

Another article for Hall of Fame entry was published in the *Journal of Quantitative Analysis in Sports*. "Using Tree Ensembles to Analyze National Baseball Hall of Fame Voting Patterns: An Application to Discrimination in BBWAA Voting," Mills and Salaga wrote about a new method to judge hall of fame entry. The Random Forest algorithm was used to look at voting patterns and use them to predict Hall of Fame entry. The Random Forest algorithm draws a number of samples from the population, with replacement. The observations that are not chosen are "out-of-bag." "…randomly sample M predictors and choose the best split among this subset of input variables (Brian Mills, 2011)." These trees grow and create a "Random Forest." Mills and Salaga used Baseball-reference.com as their source of data, and chose the most common variables judged for pitchers and hitters. The Random Forest method had a high classification rate for hall of fame entry. This method of analysis is highly complex, and out of range for my knowledge base.

CHAPTER II

THE DATA

Data was compiled from Baseball-Reference.com. The "Hall of Fame" data set and the "rejected members" data set were compiled for each player individually. For all positions, overall batting statistics were used. Since each player was voted into the Hall of Fame based on position, fielding statistics at the specific position they were voted in by were used in the analysis (instead of overall career data). Since pitchers have statistics for fielding, batting, *and* pitching, they were separated into their own data set within the "Hall of Fame" data and then inserted into the Hall of Fame data set. Overall pitching statistics were used in the analysis.

The "Rejected Members" data set was compiled the same way. If a player was listed to have played multiple positions, the position that the player played the most games at was used. Since only one position per player is being analyzed, models may have slight discrepancies from other models that may be found.

All data was compiled from Baseball-reference.com. There is a disclaimer on the website that fielding statistics may contain discrepancies. Since there may be discrepancies in the data source itself, models will be as accurate as possible based on the data given and the chosen methods described above.

The "Hall of Fame" dataset contains 83 variables and 237 players. This includes batting, fielding, and pitching variables. Since not all players have all statistics, data fields will be empty accordingly. This data set includes members of the Baseball Hall of Fame from 1936 to 2012. The "Rejected Members" data set contains 889 players and 83 variables (same as above). This list is compiled from 1936 to 2013. It includes all nominees, from 1936, and players were listed by the first time they were nominated (to ensure that all players were only listed once. All first time nominees in 2013 were also included in the dataset.

The "Hall of Fame" and "Rejected Members" datasets were merged together to create a master data set consisting of 83 variables and 1126 players. All data sets were compiled in MS Excel 2010 and read into Statistical Analysis Software (known as SAS for the remainder of this thesis). Data was copied and pasted into SPSS 19 for mean substitutions and re-saved into Excel 2010 data sheets. Each of the 9 data sets was then read into SAS for further analysis for the time periods section of the thesis. As for the Tigers and Nationals rosters for September 4, 2013, they were compiled the same way as previously stated. The rosters were found on sbnation.com. The list for the top 25 current MLB players was found on ESPN.com. Players were listed by current position, however to keep the methods the same, players were placed in the position that they played the most games at.

The following is a list of codes that were implemented into the data sets using SAS:

Hall of Fame Member:

$$\mathbf{HOF1=}\begin{cases}\mathbf{0}, & \mathbf{NO}\\ \mathbf{1}, & \mathbf{YES}\end{cases}$$

This assignment will later be used as the dependent variable in various logistic regressions

Time Period:

$$\mathbf{TIMEPERIOD=}\begin{cases}\mathbf{1}, & \mathbf{1876\ AND\ PRIOR}\\ \mathbf{2}, & \mathbf{1910-1941}\\ \mathbf{3}, & \mathbf{1941-1974}\\ \mathbf{4}, & \mathbf{1974-PRESENT}\end{cases}$$

This assignment will later be used in to split up each position into different time periods. This will be used as the sorting variable when the time period regressions are run.

Position Number:

$$\mathbf{POSITION=}\begin{cases}\mathbf{1}, & \mathbf{PITCHER}\\ \mathbf{2}, & \mathbf{CATCHER}\\ \mathbf{3}, & \mathbf{FIRST\ BASE}\\ \mathbf{4}, & \mathbf{SECOND\ BASE}\\ \mathbf{5}, & \mathbf{THIRD\ BASE}\\ \mathbf{6}, & \mathbf{SHORT\ STOP}\\ \mathbf{7}, & \mathbf{LEFT\ FIELD}\\ \mathbf{8}, & \mathbf{CENTER\ FIELD}\\ \mathbf{9}, & \mathbf{RIGHT\ FIELD}\end{cases}$$

These assignments are used to split the master data set into the nine different position

Mathematics Background

This analysis was based on logistic regression. Within logistic regression, there are many statistics being computed and probability values being found. First, here is an overview of logistic regression: what it is and how it works.

Logistic regression is used to obtain the probability that a desired outcome will occur (dependent variable) given values for many independent variables. In the case of this thesis, logistic regression was used to construct models for predicting hall of fame entry.

$$\text{HOF1}=\begin{cases}1, & yes\\ 0, & no\end{cases}$$

In logistic regression, the variable being predicted needs to be categorical, with only two categories: it happened or it did not. The main equations in logistic regression are

$$P(Y=1)=\frac{1}{1+e^{-z}} \quad (1)$$

$$\text{Where } z=\alpha+b_1X_1+b_2X_2+b_3X_3+\ldots\ldots+bnX_n \quad (2)$$

Logistic regression models the probability of the odds. The odds of an event occurring is modeled by equation 3 and is defined from zero to positive infinity.

$$\text{Odds } (Y=1)=\frac{P(Y=1)}{1-P(Y=1)} \quad (3)$$

The transformation from the odds (3) to the probability (1) can be seen using simple algebra. If above we let P be the P(Y=1), then 1-P = P(Y=0).

$$P=\frac{P}{1}$$

$$=\frac{P}{1-P+P}$$

$$=\frac{P}{(1-P)+P}$$

$$=\frac{(1-P)}{(1-P)}\cdot\frac{P}{(1-P)+P}$$

$$=\frac{\frac{P}{(1-P)}}{\frac{(1-P)}{(1-P)}+\frac{P}{(1-P)}}$$

$$=\frac{\frac{P}{(1-P)}}{1+\frac{P}{(1-P)}}$$

$$=\frac{ODDS}{1+ODDS}$$

Z=*ln(odds)*=equation 2

$$Z= B_0+B_1X_1 +BX_2+ B_3X_3 +\ldots\ldots+B_nX_n$$

$$e^{\ln(odds)} = odds$$

Substituting into equation (2),

$$= \frac{e^z}{1+e^z}$$

$=\frac{1}{1+e^{-z}}$ which is equivalent to (1).

Taking the natural log of the odds makes the function continuous from negative infinity to positive infinity. The equation is modeled below.

$$\text{Logit (Y)} = \ln\left(\frac{(P(Y=1)}{1-P(Y=1)}\right) = ln(\frac{Y}{1-Y}) \quad (4)$$

This equation can be modeled as equation (2) and the probability that an event occurs is equation (1) (both of which were mentioned earlier in this thesis). By making the logit transformation, the range is 0 to 1 to encompass a probability. The Logistic regression uses maximum likelihood to make prediction probabilities. "…The method of maximum likelihood yields values for the unknown parameters which maximize the probability of obtaining the observed set of data " (Hosmer & Lemeshow, 2000) Z is called the maximum likelihood estimator of the binomial distribution. (Hosmer & Lemeshow, 2000).

Since logistic regression deals with categorical variables with 2 outcomes, it has a Bernoulli distribution. A Bernoulli distribution has a probability mass function of

$$f(Y|x) = p^x(1-p)^{1-x}$$

The maximum likelihood estimator (MLE) is modeled below:

$$\mathbf{\Pi} p^x(1-p)^{1-x}$$

$$\theta = p^{\Sigma x}(1-p)^{n-\Sigma x}$$

$$ln(\theta) = \Sigma x ln(p) + (n - \Sigma x) ln(1 - p)$$

$$\frac{d\, ln(\theta)}{d\theta} = \frac{\Sigma x}{p} + \frac{(n - \Sigma x)}{(1 - p)}\ (-1)$$

$$0 = \frac{\Sigma x}{p} - \frac{(n - \Sigma x)}{(1 - p)}$$

$$\frac{\Sigma x}{p} = \frac{(n - \Sigma x)}{(1 - p)}$$

$$(1 - p)\Sigma x = (n - \Sigma x)p$$

$$\Sigma x - p\Sigma x = (n - \Sigma x)p$$

$$\Sigma x = (n - \Sigma x)p + p\Sigma x$$

$$\Sigma x = np - p\Sigma x + p\Sigma x$$

$$\frac{\Sigma x}{n} = p$$

Expected value= p

Any unknown parameters are estimated by the MLE, or the expected value in this case. It is the number of players that are expected to make it into the hall of fame. MLE approach can have both numerical and categorical independent variables (Kleinbaum, Klein, & Pryor, 2002).

$$P(Y = 1) = \frac{1}{1 + e^{-z}} \qquad (1)$$

Where $z = \alpha + b_1X_1 + b_2X_2 + b_3X_3 + \ldots\ldots + bnX_n$ (2)

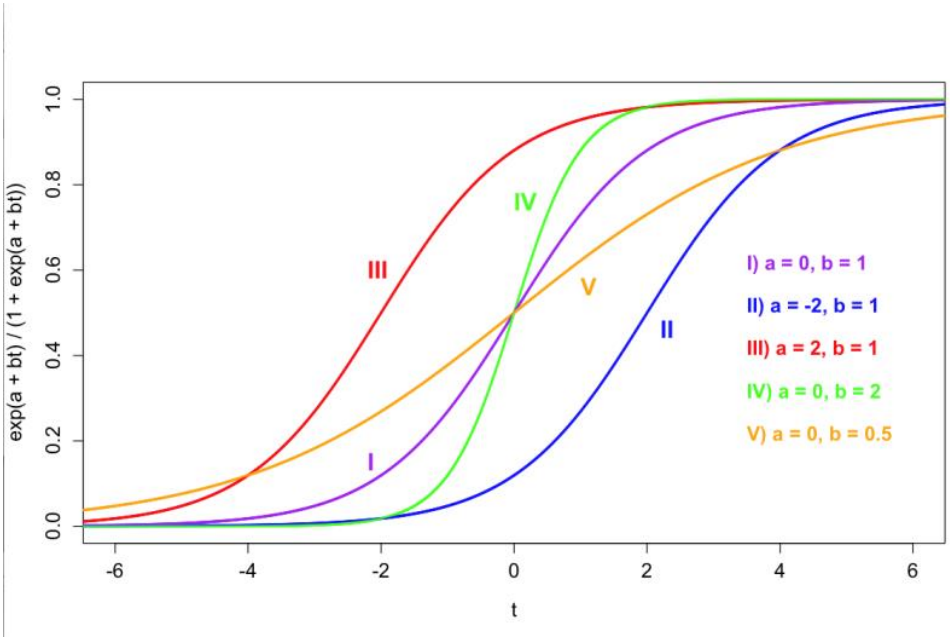

Figure 2.1: *Transformation of the Logistic Curve*

The graph depicted above is the logistic curve. It shows the different transformations that occur (Johnson D. E., 2013).

As for the tests that are being computed internally in the program, this section starts the many hypothesis tests that are taking place behind the scenes. To do this, an example of one of the regressions will be used as an example. This is a piece of the output that SAS produced for the original regression run for first basemen.

R-Square	0.4352	**Max-rescaled R-Square**	0.7980

The r-squared value describes the amount of variation explained by the model produced. A high r-squared value is desired with 0.49 being "strong" and 0.3 being "moderate".

Table 2.1

Testing Beta Weights Equal to Zero

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	46.8499	3	<.0001
Score	35.6693	3	<.0001
Wald	5.5019	3	0.1385

The first hypothesis test comes into play at this level. Below is what is happening:

$$H_o: \text{Betas} = 0$$

$$H_a: \text{At least one of the Betas} \neq 0$$

$$P=0.05$$

$$\chi^2 = 5.5019 \text{ df}=3 \text{ p-value}=0.1385$$

A chi-squared statistic is used here to see if all betas are equal to zero or if at least one differs. Below is the formula for a chi-squared statistic.

$$\chi^2=\Sigma \frac{(\text{observed}-\text{expected})^2}{\text{expected}} \quad (5)$$

where *expected*= $\frac{(row\ total)(column\ total)}{table\ total}$ (6)

and *degrees of freedom = (number of rows – 1)*(number of columns – 1)* (7)

Since p>0.05, the null hypothesis is rejected. At least one Beta is not equal to zero.

Residual Chi-Square Test		
Chi-Square	**DF**	**Pr > ChiSq**
19.6454	27	0.8453

This tests the hypothesis that:

H_o: The residuals indicate a good "goodness of fit"

H_a: The residuals do not indicate a good "goodness of fit"

P=0.05

χ^2= 19.6454 df=27 p-value=0.8453

Since our p-value is greater than 0.05, the null hypothesis is not rejected. Therefore the residuals indicate a good "goodness of fit."

Table 2.2

Summary of Stepwise Selection

Summary of Stepwise Selection								
Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > Chi Sq	Variable Label
	Entered	Removed						
1	SLUGGING__		1	1	18.6613		<.0001	SLUGGING %
2	TRIPLES		1	2	12.4554		0.0004	TRIPLES
3	RBIS		1	3	9.8443		0.0017	RBIS
4	CAUGHT_STEALING		1	4	4.1628		0.0413	CAUGHT STEALING
5		SLUGGING__	1	3		3.0488	0.0808	SLUGGING %
6	PLATE_APPEARANCES		1	4	5.6501		0.0175	PLATE APPEARANCES
7		CAUGHT_STEALING	1	3		0.9872	0.3204	CAUGHT STEALING

This table shows the variables that were deemed important in the stepwise procedure. According to the SPSS survival manual, the stepwise method allows the researcher to "specify a large group of potential predictors and the program picks a subset that provides the best predictive power." Above are the group of predictors that SAS chose as the best predictors. The model was tested stepping in these variables. Once the Wald statistic

becomes “insignificant” (p-value > 0.05), a variable needs to be “stepped out.” In the example above, slugging percent and caught stealing were thrown out of the model.

Table 2.3

Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-18.9221	7.9195	5.7088	0.0169
PLATE_APPEARANCES	1	-0.00379	0.00186	4.1353	0.0420
TRIPLES	1	0.0975	0.0439	4.9359	0.0263
RBIS	1	0.0334	0.0152	4.8505	0.0276

Many hypothesis tests are becoming computed here; one for each variable chosen. As an example, plate appearances are going to be tested.

H_o: Plate appearances contribute significantly to the model to predict Hall of Fame entry.

H_a: Plate appearances do not contribute significantly to the model to predict Hall of Fame entry.

P=0.05

χ^2= 5.7088 df=1 p-value=0.0420

Since p <0.05, the null hypothesis is rejected. Plate appearances contribute significantly to the predictive model. Again, this test is completed for each variable. In the table above, all p- values are less than 0.05. Therefore, once all hypotheses are tested, all variables are said to be important to the prediction model. The values in the "Estimate" column are the coefficients to those variables in the model. If values are positive, the variables contribute positively to the model, increasing probability. If values are negative, it decreases probability. The model is given below:

$$Y= -0.00379PlateAppearances+0.0975Triples+0.0334RBIS-18.9221$$

The following is an interpretation: A 1 appearance increase in the number of plate appearances decreases the probability of Hall of Fame entry by 0.379 percent, all other variables held constant.

Table 2.4

Odds Ratio Estimates

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
PLATE_APPEARANCES	0.996	0.993	1.000
TRIPLES	1.102	1.012	1.201
RBIS	1.034	1.004	1.065

For every plate appearance, the odds of making it into the Hall of Fame decreases by 0.996, all other factors being equal.

A confidence interval is an interval that contains the true value of a number in repeated sampling. In 95 % of intervals constructed, the true odds ratio will fall within the range described. The confidence interval says that "I am 95% confident" that the true value of the odds ratio is between 0.993 and 1."

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
1.5557	7	0.9803

Hosmer-Lemeshow tests the following hypothesis:

H_o: The model is a good fit for the data.

H_a: The model is not a good fit for the data.

P=0.05

χ^2= 1.5557 df=7 p-value=0.9803

Since p> 0.05, the null hypothesis is not rejected. Therefore, the model is deemed a good fit for the data.

The last item looked at is the classification percent. This was the number of correctly identified outcomes (a player not in the hall and predicted to not be in and a player in the hall predicted to be in). SAS produces output of this, however, percent correctly classified was calculated by MS Excel. A high classification percent was the goal.

Table 2.5

List of Variables

Name	Variable	Description
PITCHING STATISTICS		
Wins	W	Number of Wins
Losses	L	Number of Losses
Win-Loss %	W_L__	$\frac{w}{w+L}$
EARNED RUNS ALLOWED	ERA	$\frac{9 * EARNED\ RUNS}{INNINGS\ PITCHED}$
GAMES	G_1	NUMBER OF GAMES PLAYED
GAMES STARTED	GS_1	NUMBER OF GAMES STARTED
GAMES FINISHED	GF	NUMBER OF GAMES FINISHED
COMPLETE GAMES	CG_1	NUMBER OF COMPLETE GAMES PITCHED
SHUT OUTS	SHO	NO RUNS ALLOWED AND A COMPLETE GAME
SAVES	SV	NUMBER OF SAVES
INNINGS PITCHED	IP	$\frac{9 * EARNED\ RUNS}{INNINGS\ PITCHED}$
HITS	H	NUMBER OF HITS/HITS ALLOWED
RUNS	R	NUMBER OF RUNS/RUNS ALLOWED
EARNED RUNS	ER	NUMBER OF EARNED RUNS ALLOWED
HOMERUNS	HR	NUMBER OF HOMERUNS HIT/ALLOWED
WALKS	BB	NUMBER OF WALKS

(Continued)

INTENTIONAL WALKS	IBB	NUMBER OF INTENTIONAL WALKS
STRIKE OUTS	SO	NUMBER OF STRIKE OUTS
HIT BY PITCH	HBP	TIMES HIT A PLAYER WITH THE PITCH
BALKS	BK	NUMBER OF BALKS
WILD PITCHES	WP_1	NUMBER OF WILD PITCHES
BATTERS FACED	BF	NUMBER OF BATTERS FACED
EARNED RUNS ALLOWED PLUS	ERA__	$\left(\frac{LEAGUE\ ERA}{ERA}\right)$*100
WHIP	WHIP	$\frac{WALKS + HITS}{INNINGS\ PITCHED}$
HITS PER 9 INNINGS	H_9	$\frac{9 * HITS}{INNINGS\ PITCHED}$
HOMERUNS PER 9 INNINGS	HR_9	$\frac{9 * HOMERUNS}{INNINGS\ PITCHED}$
WALKS PER 9 INNINGS	BB_9	$\frac{9 * WALKS}{INNINGS\ PITCHED}$
STRIKE OUTS PER 9 INNINGS	SO_9	$\frac{9 * STRIKEOUTS}{INNINGS\ PITCHED}$
STRIKE OUTS PER WALK	SO_BB	$\frac{STRIKE\ OUTS}{WALKS}$
BATTING STATISTICS		
GAMES	GAMES_PLAYED	NUMBER OF GAMES PLAYED
PLATE APPEARANCES	PLATE_APPEARANCES	NUMBER OF TIMES A PLAYER IS AT THE PLATE
AT BATS	AT_BATS	NUMBER OF AT BATS
RUNS	RUNS	NUMBER OF RUNS
HITS	HITS	NUMBER OF HITS
DOUBLES	DOUBLES	NUMBER OF DOUBLES
TRIPLES	TRIPLES	NUMBER OF TRIPLES
HOMERUNS	HOME_RUNS	NUMBER OF HOME RUNS
RUNS BATTED IN	RBIS	NUMBER OF RUNS BATTED IN
STOLEN BASES	STOLEN_BASES	NUMBER OF STOLEN BASES

(Continued)

CAUGHT STEALING	CAUGHT_STEALING	NUMBER OF TIME CAUGHT STEALING
WALKS	WALKS	NUMBER OF WALKS
STRIKE OUTS	STRIKE_OUTS	NUMBER OF STRIKE OUTS
BATTING AVERAGE	BATTING_AVERAGE	$\frac{HITS}{AT\ BATS}$
ON BASE PERCENT	ON_BASE__	$\frac{\boldsymbol{HITS + WALKS + HIT\ BY\ PITCH}}{\boldsymbol{AT\ BATS + WALKS + HIT\ BY\ PITCH + SACRIFIC}}$
SLUGGING PERCENT	SLUGGING__	$\frac{TOTAL\ BASES}{AT\ BATS}$
ON BASE PLUS SLUGGING PERCENT	OPS	ON BASE+SLUGGING PERCENTS
ADJUSTED ON BASE PERCENT	OPS_	$100 * [\frac{\left(\frac{OBP}{LG\ OBP}\right)}{\left(\frac{SLG}{LG\ SLG - 1}\right)}]$
TOTAL BASES	TOTAL_BASES	SINGLES+2*DOUBLES +3*TRIPLES+4*HOME RUNS
DOUBLE PLAYED GROUND INTO	GDP	NUMBER OF DOUBLE PLAYS GROUNDED INTO
HIT BY PITCH	HBP	NUMBER OF TIMES HIT BY PITCH
SACRIFICE HITS	SACRIFICE_HITS	NUMBER OF SACRIFICE HITS
SACRIFICE FLIES	SACRIFICE_FLY	NUMBER OF SACRIFICE FLIES
INTENTIONAL WALKS	INTENTIONAL_WALKS	NUMBER OF TIMES INTENTIONALLY WALKED
FIELDING STATISTICS		
GAMES	G	NUMBER OF GAMES PLAYED
GAMES STARTED	GS	NUMBER OF GAMES STARTED
COMPLETE GAMES	CG	NUMBER OF COMPLETE GAMES PLAYED
INNINGS PLAYED	Inn	NUMBER OF INNINGS PLAYED
DEFENSIVE CHANCES	Ch	PUTOUTS+ASSISTS+ERRORS
PUTOUTS	PO	NUMBER OF PUTOUTS
ASSISTS	A	NUMBER OF ASSISTS

(Continued)

ERRORS	E	NUMBER OF ERRORS COMMITTED
DOUBLE PLAYS TURNED	DP	NUMBER OF DOUBLE PLAYS TURNED
FIELDING PERCENT	FLD__	$\frac{PUTOUTS + ASSISTS}{PUTOUTS + ASSISTS + ERRORS}$
TOTAL FIELDING RUNS ABOVE AVERAGE	RTOT	TOTAL FIELDING RUNS ABOVE AVERAGE
TOTAL FIELDING RUNS ABOVE AVERAGE PER YEAR	RTOT_YR	TOTAL FIELDING RUNS ABOVE AVERAGE PER 1200 INNINGS
DEFENSIVE RUNS SAVED ABOVE AVERAGE	RDRS	DEFENSIVE RUNS SAVED ABOVE AVERAGE
DEFENSIVE RUNS SAVED ABOVE AVERAGE PER YEAR	RDRS_YR	DEFENSIVE RUNS SAVED ABOVE AVERAGE PER 1200 INNINGS
LEAGUE FIELDING PERCENT	LgFld_	FIELDING PERCENT FOR THE LEAGUE AT THIS TIME
LEAGUE RANGE FACTOR PER 9 INNINGS	LgRF9	AVERAGE RANGE FACTOR PER 9 INNINGS FOR THE LEAGUE
LEAGUE RANGER FACTOR PER GAME	LgRFG	AVERAGE RANGE FACTOR PER GAME FOR THE LEAGUE
STOLEN BASES	STOLEN_BASES	NUMBER OF STOLEN BASES
CAUGHT STEALING	CAUGHT_STEALING	NUMBER OF TIMES CAUGHT STEALING
CAUGHT STEALING	CS	$\frac{CAUGHT\ STEALING}{CAUGHT\ STEALING+STOLEN\ BASES}$

(Continued)

PERCENT		
LEAGUE CAUGHT STEALING PERCENT	CS__	$\frac{\text{LEAGUE EXPECTED CAUGHT STEALING}}{\text{PLAYERS CAUGHT STEALING + PLAYERS STOLEN B}}$
PICK OFFS	PO_1	NUMBER OF TIMES PICKED OFF
PASSED BALLS	PB	NUMBER OF PASSED BALLS
WILD PITCHES	WP	NUMBER OF WILD PITCHES

CHAPTER III

RESEARCH METHODS

The first step in running the analysis was to create a prediction model for future Hall of Fame members. This was obtained by running logistic regressions in SAS for the nine different positions using a stepwise method. First, the data was read into SAS and split into 9 different datasets: one for every position. The data was examined for missing values in variables. The variables that had the majority of data were used in the logistic regression as independent variables to predict Hall of Fame. A logistic regression was then run for each position and a model was obtained. Examining and interpreting chi-squared statistics, pseudo- r^2 values, and classification tables provided the first set of models. Hosmer-Lemeshow statistics and p-values were used to test goodness-of-fit, however, these statistics were used as a secondary way of showing the accuracy of a model. Models in which the Hosmer-Lemeshow p-values were not large enough to prove goodness-of-fit were re-run using the enter method and only the variables that were picked as predictors in the stepwise regression before were used.

Another logistic regression was performed for each position, this time with additional variables that had more missing values. This was done to see if a model would be more accurate with the addition of more variables. Once the models are determined, the r^2 values and classification tables were compared to the first set of models.

Another logistic regression was run to compare to the other sets of models. This time, the missing values were filled in with the means of those variables. This was done using SPSS series mean substitutions for each variable after the main data set was split into the different positions. The reasoning was that players of the same position will likely have similar statistics. If the models are similar, then the new model was also considered. If the new model sets were completely different, the model cannot be used and a previous model will be final. To see if other variables could potentially be used as predictors, a final logistic regression was performed using all variables (pitching variables were only included in the pitching regression). Again, the models were created and r^2 values and classification tables were used in the analysis.

After running all of the regressions, a check to make sure the models were good fit was completed. This was done by looking at r^2 values and classification tables. Since not all of the models had a high enough Hosmer- Lemeshow p-value, but good classification percentages, classification tables were used as the primary source of goodness-of-fit. Hosmer-Lemeshow was used as a secondary source.

Another goal of the analysis was to compare the positions across different time periods. To do this, the master data set was then split into different time eras based on ball construction. Four time periods were established in which major changes to the ball occurred. The players were then split up into different time frames based on the majority of years played in each time period. A person that played an equal amount of time in

different time frames was given the latter time period number. Logistic regression methods were done as described above and, again, models were chosen based on r^2 values and classification tables. Models were then compared between the time eras and differences were determined.

For all models (for time period analysis as well as overall position analysis) classification percentages were calculated in MS Excel 2010. Using the formulas

$$P(Y = 1) = \frac{1}{1+e^{-z}} \quad (1)$$

$$\text{Where } z = \alpha + b_1\mathbf{X_1} + b_2\mathbf{X_2} + b_3\mathbf{X_3} + \ldots\ldots + bn\mathbf{X_n} \quad (2)$$

Probabilities were calculated using equation 1. (Information about these formulas can be found in appendix one.) If a player did not make it into the Hall of Fame and had less than 50 percent as their probability, this was considered correctly classified. If a player made it into the Hall of Fame and had a probability greater than 50 percent, this was considered correctly classified. From here, simple percentages were computed to get the overall correct classification percent that was used in determining which model best fit the data.

Once all models have been obtained, predictions can take place. The first set of predictions will be for making it into the Hall of Fame. Using the members of the starting lineups for the Detroit Tigers and the Washington Nationals on September 4, 2013, players were put into the models and the probability that the players on these teams make

it into the Hall of Fame were computed. Also a list of the top 25 best players in Major League Baseball (MLB) was compiled and run through the same analysis. If a player's probability of making it into the hall is 0.5 or greater, then this player will be said to have a decent shot at making it in. If a player's probability is greater than 0.75, then the player is said to have a good shot of making it in. Any probability that is 0.90 or higher will be considered a "shoe in."

Using the same concepts as presented above, two members from each time frame that are in the Hall of Fame were selected. The members from the other time eras were then be put into the regression equations found and it was determined if they would make it into the Hall of Fame within the other time periods, assuming models exist.

CHAPTER IV

RESULTS

The regressions shown are the ones that were chosen. A summary table of all regressions can be found in appendix A: Additional Regression Information. The following are the important pieces of the analysis for models chosen. Information about the mathematics that is being computed by the program can be found in chapter two. The tables listed are the models chosen for position regression, as well as time period analysis in the latter. Interpretation follows.

R-squared values are interpreted as a percent of total variation that is explained the by model found. It reduces initial uncertainty by the percent number. Testing global null hypothesis that all beta weights are equal to zero, the Wald statistic is what was analyzed. However, if the Wald statistic was not shown to be important, the likelihood ratio and score were looked at. The residual chi-squared test shows the relationship between the residuals of the model and the independent variables selected in the step wise procedure. It is another way to show goodness of fit. If the p-value is greater than 0.05, it is concluded that the model is a good fit. SAS picked out variables as predictors that were deemed significant. Beta weights are also tested to see if they are important contributors. The column labeled “estimate” shows the beta weights; these are the coefficients to the contributing variables in the model equation. In order to obtain probabilities, the resulting

number from the equation needs to be plugged in to equation one and evaluated. Hosmer-Lemeshow tests the overall goodness of fit. As previously mentioned, the Hosmer-Lemeshow test was taken as a secondary goodness of fit test; r-squared values, residual tests, and classification percentages were taken as the primary goodness of fit determination.

The table below lists all of the models chosen at each position. Also listed are all of the important tests and results run in SAS. Not all numbers and statistics were interpreted in output; only the main tests were used. Models listed use the names of the variables in the analysis, not the names of the statistic. These names, as well as their meaning and calculation methods, can be found in Chapter two.

Table 4.1

Regression Analysis per Position

REGRESSION ANALYSIS PER POSITION										
		GLOBAL NULL: BETA=0								
		LIKELIHOOD RATIO			SCORE			WALD		
	R^2	X^2	DF	P-VALUE	X^2	DF	P-VALUE	X^2	DF	P-VALUE
PITCHER	0.3988	144.508	4	<0.0001	127.3009	4	<0.0001	32.8269	4	<0.0001
CATCHER	0.2521	15.3964	1	<0.0001	14.7299	1	<0.0001	8.9577	1	0.0028
FIRST BASE	0.4352	46.8499	3	<0.0001	35.6693	3	<0.0001	5.5019	3	0.1385
SECOND BASE	0.5071	41.0288	3	<0.0001	31.8092	3	<0.0001	5.9157	3	0.1158
THIRD BASE	0.3922	41.3312	2	<0.0001	31.3439	2	<0.0001	11.1617	2	0.0038
SHORT STOP	0.5053	49.2692	2	<0.0001	34.6049	2	<0.0001	12.0104	2	0.0025
LEFT FIELD	0.3356	37.6149	1	<0.0001	33.9059	1	<0.0001	19.3368	1	<0.0001
CENTER FIELD	0.5227	56.9545	2	<0.0001	43.7554	2	<0.0001	7.9251	2	0.019
RIGHT FIELD	0.539	66.593	2	<0.0001	49.8408	2	<0.0001	9.4574	2	0.0088

(Continued)

REGRESSION ANALYSIS PER POSITION CONTINUED					
	RESIDUAL CHI-SQUARE TEST			SUMMARY OF STEPWISE SELECTION	
	X^2	DF	P-VALUE	ENTERED	REMOVED
PITCHER	51.5676	56	1	CG SV WHIP W_L_	
CATCHER	31.2552	30	0	RUNS WP	WP
FIRST BASE	19.6454	27	1	PLATE_APPEARANCES TRIPLES RBIS	
SECOND BASE	13.8776	27	1	RBIS STRIKE_OUTS GDP	
THIRD BASE	25.7265	25	0	HITS YRS RF_G	RF_G
SHORT STOP	24.8158	27	1	RBIS PO STRIKE_OUTS	STRIKE_OUTS
LEFT FIELD				RUNS	
CENTER FIELD	23.5638	25	1	OPS E DP	DP
RIGHT FIELD	67.1389	28	<0.0001	TRIPLES RUNS	

(Contiued)

REGRESSION ANALYSIS PER POSITION CONTINUED						
POSITION	MAXIMUM LIKELIHOOD ANALYSIS					
	PARAMETER	DF	ESTIMATE	STANDARD ERROR	WALD CHI SQUARED	P-VALUE
	INTERCEPT		4.2170	6.8118	0.3825	0.5363
	CG	1	0.0423	0.00772	30.0369	<0.0001
	W_L_	1	17.6976	7.3612	5.78	0.0162
	SV	1	0.0223	0.00525	18.0388	<0.0001
PITCHER	WHIP	1	-18.3097	4.5711	16.0441	<0.0001
	INTERCEPT	1	-7.1944	2.0825	11.9346	0.0006
CATCHER	RUNS	1	0.00696	0.00233	8.9577	0.0028
	INTERCEPT	1	-18.9221	7.9195	5.7088	0.0169
	PLATE_APPEARANCES	1	-0.00379	0.00186	4.1353	0.0420
	TRIPLES	1	0.0975	0.0439	4.9359	0.0263
FIRST BASE	RBIS	1	0.0034	0.0152	4.8505	0.0276
	INTERCEPT	1	-19.5313	8.2756	5.5701	0.0183
	RBIS	1	0.0236	0.0973	5.8834	0.0153
	GDP	1	0.0347	0.0175	3.9353	0.0473
SECOND BASE	STRIKE_OUTS	1	-0.00924	0.00412	5.0174	0.0251
	INTERCEPT	1	-4.8687	2.6035	3.4971	0.0615
	YRS	1	-1.0770	0.3810	7.9892	0.0047
THIRD BASE	HITS	1	0.0101	0.00305	10.9395	0.0009
	RBIS	1	0.00905	0.00303	8.8908	0.0029
	PO	1	0.00202	0.000771	6.8855	0.0087
SHORT STOP	INTERCEPT	1	-14.5948	4.1031	12.6523	0.0004
	INTERCEPT	1	-6.1367	1.2259	25.0592	<0.0001
LEFT FIELD	RUNS	1	0.00455	0.00103	19.3368	<0.0001
	INTERCEPT	1	-55.6474	21.3350	6.8031	0.0091
	OPS	1	58.8519	23.8337	6.0973	0.0135
CENTER FIELD	E	1	0.0836	0.0310	7.2492	0.0071
	INTERCEPT	1	-17.3814	5.4559	10.1491	0.0014
	TRIPLES	1	0.00893	0.00303	8.6754	0.0032
RIGHT FIELD	RUNS	1	0.0632	0.0266	5.6580	0.0174

(Continued)

REGRESSION ANALYSIS PER POSITION CONTINUED						
POSITION	**MODEL**	**ODDS RATIO**		**HOSMER-LEMESHOW**		
		EFFECT	**POINT ESTIMATE**	χ^2	**DF**	**P-VALUE**
PITCHER	Y=0.0423**CG**+17.6976**W_L_**+0.0223**SV**-18.3097**WHIP**+4.1270	CG W_L_ SV WHIP	1.043 >999.999 1.023 <0.001	202.9661	8	<0.0001
CATCHER	Y=0.00696**RUNS**-7.1944	RUNS	1.007	6.6717	9	0.6713
FIRST BASE	Y=-.00379**PLATE_APPEARANCES**+0.0975**TRIPLES**+0.0334**RBIS**-18.9221	PLATE_APPEARANCES TRIPLES RBIS	0.996 1.102 1.034	1.5557	7	0.9803
SECOND BASE	Y=0.0236**RBIS**+0.0347**GDP**-0.00924**STRIKEOUTS**-19.5313	RBIS GDP STRIKE_OUTS	1.024 1.035 0.991	1.581	7	0.9794
THIRD BASE	Y=-1.0770**YEARS**+0.0101**HITS**-4.8687	YRS HITS	0.341 1.010	2.934	8	0.9384
SHORT STOP	Y=0.00905**RBIS**+0.00202**PO**-14.5948	RBIS PO	1.009 1.002	3.6533	8	0.887
LEFT FIELD	Y=0.00455**RUNS**-6.1367	RUNS	1.005	23.4475	8	0.0028
CENTER FIELD	Y=58.8519**OPS**+0.0836**E**-55.6474	OPS E	>999.999 1.087	1.4151	8	0.994
RIGHT FIELD	Y=0.00893**RUNS**+0.0632**TRIPLES**-17.3814	RUNS TRIPLES	1.009 1.065	9.1616	7	0.2413

By Position

Pitchers

The r-squared value is moderate. 39.88% of total variation is explained by the set of variables. For every complete game, the odds of making it into the Hall of Fame increase by a factor of 1.043, all other factors being equal. For every win-loss percent, the odds of making it into the Hall of Fame increase by a factor of >999, all other factors being equal. For every save, the odds of making it into the Hall of Fame increase by a factor of 1.023, all other factors being equal. For every WHIP, the odds of making it into the Hall of Fame increase by a factor of <0.01, all other factors being equal. The model given fails the Hosmer-Lemeshow goodness of fit test. Since the r-squared value is moderate and the residual goodness of fit is good, the classification tables were taken into consideration for model fit. After analyzing all of the classifications, the correct classification percent is 91.32%. Therefore, this model was deemed a good fit.

Catchers

The r-squared value is moderate. 25.21 % of variation is explained by the set of independent variables chosen by this model. For every additional run, the odds of making it into the Hall of Fame increases by a factor of 1.007. Hosmer-Lemeshow: this is a secondary resource to test for goodness of fit. Since 0.6713>0.05, the model is deemed good. However, correct classification percentages were looked at as greater indication. 86.87% of players were correctly classified.

First Basemen

R-squared value is strong. This is a good sign! 43.53% of variation is explained by the set of variables found in this model as predictors.

Testing the global beta = 0: The Wald statistic is questionable. This is when the other items are taken into consideration. Since the other tests look good, the analysis was continued.

The residual test looks good. The model is a good fit by the residual test. For every plate appearance, the odds of making it into the Hall of Fame decrease by a factor of 0.996. For every triple, the odds of making it into the Hall of Fame increase by a factor of 1.102. For every RBI, the odds of making it into the Hall of Fame increase by a factor of 1.034. Hosmer-Lemeshow criteria met for goodness of fit. Correct classification was 90.1%. Therefore, the model is deemed good.

Second Basemen

It is a strong r-squared value. 50.71 % of variation is explained by the variables chosen in this model. The Wald statistic is questionable for the global null hypothesis that beta = 0. However, the other tests look good, so the model is considered. The residual test for goodness of fit came out good! No variables were removed in the stepwise summary. All were considered significant contributors. For every RBI, the odds of making it into the Hall of Fame increase by a factor of 1.024, all other factors being equal. For every double play grounded in to, the odds of making it into the Hall of Fame increase by a factor of 1.035, all other factors being equal. For every strike out, the odds of making it into the Hall of Fame decrease by a factor of 0.991, all other factors being equal. The Hosmer-Lemeshow goodness of fit test is good implying the model is a good fit for the data. The correct classification percent is 93.98 %.

Third Basemen

The r-squared value is moderate. 39.22% of variation is explained by the variables selected in this model. The Wald test for global null hypothesis that betas = 0 came out good with $p<0.05$. The residual goodness of fit test looks good with $p>0.05$.RF_G was removed because the model was deemed insignificant with it included (p-value=0.0992). All other p values are less than 0.05. Therefore all of the other listed variables remain in the model. For every year played, the odds of making it into the Hall of Fame decrease by a factor of 0.341, all other factors being equal. For every hit, the odds of making it into the Hall of Fame increase by a factor of 1.010, all other factors being equal. The Hosmer-Lemeshow test turned out to support the idea of goodness of fit for the model. The correct classification was 85.39%.

Shortstop

The r-squared value is strong. 50.53% of variation is explained by the variables found in this model. The Wald statistic had a significant p-value associated with it therefore supporting the analysis.The residual test's p-value>0.05 indicating that the model is a good fit for the data.

Strike outs were removed because criteria were not met. The p-values indicate that all other variables are found to be predictors. For every RBI, the odds of making it into the Hall of Fame increase by a factor of 1.009, all other factors being equal. For every Pick off, the odds of making it into the Hall of Fame increase by a factor of 1.002, all other factors being equal. The Hosmer-Lemeshow also supports the goodness of fit for the model. The correct classification is 87.06%.

Left Field

The r-squared value is moderate. 33.56% of variation is explained by variables identified by this model. The Wald test that not all weights are zero came out good! Therefore, the analysis can continue.

Runs are the only significant contributor to left fielders. For every run, the odds of making it into the Hall of Fame increase by a factor of 1.005. The Hosmer-Lemeshow goodness of fit test is not satisfied. However the correct classification is 89.13%. Therefore, the model is a good model.

Center Fielder

It is a strong r-squared value. 52.27% of variation is explained by the variables in this model.

All tests had p-values <0.05. Therefore, analysis can continue because not all weights are zero.

The residual test for goodness of fit turned out good! This is another indicator that the model is a good fit of the data. DP did not make the cut to be kept in the model. All variables left in the model are important for prediction. For every OPS, the odds of making it into the Hall of Fame increase by a factor of >999, all other factors being equal. For every error, the odds of making it into the Hall of Fame increase by a factor of 1.087, all other factors being equal. Hosmer-Lemeshow also supports the model being good! The correct classification percent is 87.1%.

Right Fielder

It is a strong r-squared value. 53.90% of variation is explained by the variables selected in the model. All tests that weights are not all zero came out good! Residual test does not support the model being a good fit. Analysis is continued and other methods for testing goodness of fit will

be executed. All variables passed the Wald test to be kept in the model, and all variables are deemed important in predicting hall of fame entry for right fielders. For every run, the odds of making it into the Hall of Fame increase by a factor of 1.009, all other factors being equal. For every triple, the odds of making it into the Hall of Fame increase by a factor of 1.065, all other factors being equal. Hosmer-Lemeshow supports the model fitting the data. The correct classification is 94.68%.

These are the models chosen based largely on r-squared values, residual tests, and classification percentages. The next sets of regressions chosen were for time period analysis. When working on this, many time periods for various positions did not have regression models. Keeping this in mind, only the regressions that turned out to have models are displayed in this section. All models are presented in Appendix A.

Table 4.2

Time Period Regression Analysis

TIME PERIOD REGRESSION ANALYSIS											
			GLOBAL NULL: BETA=0								
			LIKELIHOOD RATIO			SCORE			WALD		
POSITION	TIME PERIOD	R-SQUARED	CHI SQUARED	DF	P-VALUE	CHI SQUARED	DF	P-VALUE	CHI SQUARED	DF	P-VALUE
PITCHER	1	0.5636	26.5344	1	<0.0001	18.8616	1	<0.0001	7.6414	1	0.0057
	2	0.5268	37.4166	2	<0.0001	24.3456	2	<0.0001	10.2577	2	0.0059
	3	0.4576	74.0223	4	<0.0001	58.1324	4	<0.0001	12.3296	4	0.0151
	4	0.3977	68.9521	3	<0.0001	68.3249	3	<0.0001	6.8283	3	0.0776
CATCHER	3	0.3957	19.1381	1	<0.0001	13.5687	1	0.0002	6.0011	1	0.0143
FIRST BASE	3	0.5579	29.3808	1	<0.0001	13.9883	1	0.0002	4.0581	1	0.044
SECOND BASE	3	0.4711	19.7456	1	<0.0001	17.2353	1	<0.0001	6.179	1	0.0129
THIRD BASE	NONE										
SHORT STOP	3	0.3568	17.2135	1	<0.00001	15.516	1	<0.0001	7.9886	1	0.0047
LEFT FIELD	3	0.4151	12.5537	1	<0.0001	13.5508	1	0.0002	6.4468	1	0.0111
CENTER FIELD	3	0.5321	23.5458	1	<0.0001	19.4952	1	<0.0001	4.9604	1	0.0259
RIGHT FIELD	4	0.2844	14.0563	1	0.0002	13.5	1	0.0002	5.0816	1	0.0242

(Continued)

TIME PERIOD REGRESSION ANALYSIS					
	RESIDUAL CHI-SQUARE TEST			SUMMARY OF STEPWISE SELECTION	
POSITION	**CHI SQUARED**	**DF**	**P-VALUE**	**ENTERED**	**REMOVED**
PITCHER	19.4421	30	0.9303	SO ERA_	ERA_
	26.4724	47	0.9932	SO SACRIFICE_HITS WALKS	WALKS
	59.7871	59	0.4469	SO ERA_ HOME_RUNS Fld_ YRS	YRS
	15.4371	58	1	CG_1 SV WHIP GF	SV
CATCHER	24.1101	31	0.806	RBIS CG OPS A	RBIS CG A
FIRST BASE	11.5453	30	0.999	SLUGGING__ TRIPLES	TRIPLES
SECOND BASE	19.0327	29	0.9204	SLUGGING__ Fld_	Fld_
THIRD BASE					
SHORT STOP	31.1141	32	0.5112	TRIPLES Fld_	Fld_
LEFT FIELD	56.6798	28	0.0011	RUNS	
CENTER FIELD	21.4992	29	0.8403	OPS PO	PO
RIGHT FIELD	11.4132	36	1	INTENTIONAL_WALKS STOLEN_BASES	STOLEN_BASES

(Continued)

TIME PERIOD REGRESSION ANALYSIS CONTINUED							
POSITION	**TIME PERIOD**	**MAXIMUM LIKELIHOOD ANALYSIS**					
		PARAMETER	**DF**	**ESTIMATE**	**STANDARD ERROR**	**WALD CHI SQUARED**	**P-VALUE**
PITCHER	1	INTERCEPT SO	1 1	-7.12 0.00563	2.4533 0.00204	8.2430 7.6414	0.0037 0.0057
	2	INTERCEPT SACRIFICE_HITS SO	1 1 1	-12.9187 0.0521 0.00745	3.9635 0.0248 0.00257	10.6240 4.4044 8.3838	0.0011 0.0358 0.0038
	3	INTERCEPT HOME_RUNS Fld_ SO ERA_	1 1 1 1 1	-212.2 0.1272 180.4 0.00585 0.2211	78.9478 0.0546 72.9387 0.00188 0.0760	7.2235 5.4312 6.1173 9.6589 8.4502	0.0072 0.0198 0.0134 0.0019 0.0037
	4	INTERCEPT GF CG_1 WHIP	1 1 1 1	30.9522 0.0214 0.0920 -38.7818	17.0993 0.00877 0.0381 17.0351	3.2766 5.9596 5.8371 5.1828	0.0703 0.0146 0.0157 0.0228
CATCHER	3	INTERCEPT OPS	1 1	-33.2737 40.9348	13.4375 16.7101	6.1315 6.0011	0.0133 0.0143
FIRST BASE	3	INTERCEPT TRIPLES	1 1	-45.9493 90.1684	22.5682 44.7604	4.1454 4.0581	0.0417 0.0440
SECOND BASE	3	INTERCEPT SLUGGING__	1 1	-22.6661 54.8977	8.7656 22.0849	6.6864 6.1790	0.0097 0.0129
THIRD BASE	NONE						
SHORT STOP	3	INTERCEPT TRIPLES	1 1	-6.4497 0.0797	2.0619 0.0282	9.7849 7.9886	0.0018 0.0047
LEFT FIELD	3	INTERCEPT RUNS	1 1	-5.7021 0.00493	1.9961 0.00194	8.1601 6.4468	0.0043 0.0111
CENTER FIELD	3	INTERCEPT PO	1 1	-36.4753 43.0637	15.9037 19.3353	5.2602 4.9604	0.0218 0.0259
RIGHT FIELD	4	INTERCEPT INTENTIONAL_WALKS	1 1	-9.4257 0.0555	3.8959 0.0246	5.8534 5.0816	0.0155 0.0242

(Continued

TIME PERIOD REGRESSION ANALYSIS CONTINUED						
POSITION	**MODEL**	**ODDS RATIO**		**HOSMER-LEMESHOW**		
		EFFECT	**POINT ESTIMATE**	**CHI SQUARE**	**DF**	**P-VALUE**
PITCHER	Y=0.00563**SO**-7.12	SO	1.006	2.3876	9	0.9838
	Y=0.0521**SACRIFICEHITS**+ 0.00745**SO**-12.9187	SACRIFICE_HITS SO	1.054 1.007	4.6006	8	0.7993
	Y=0.1272**HOMERUNS**+180 .4**FIELDPERCENT**+0.00585 **SO**+0.2211**ERAPLUS**-212.2	HOME_RUNS Fld_ SO ERA_	1.136 >999.999 1.006 1.247	4.378	8	0.8215
	Y=0.0214**GAMESFINISHED** +0.092**CG_1**- 38.7818**WHIP**+30.9522	GF CG_1 WHIP	1.022 1.096 <0.001	1.0804	6	0.9824
CATCHER	Y=40.9348**OPS**-33.2737	OPS	>999.999	5.0881	8	0.7481
FIRST BASE	Y=90.1684**SLUGGINGPERCENT**-45.9493	SLUGGING__	>999.999	7.351	5	0.1958
SECOND BASE	Y=54.8977**SLUGGINGPERCENT**-22.6661	SLUGGING__	<999.999	2.4593	7	0.9301
THIRD BASE						
SHORT STOP	Y=0.0797**TRIPLES**-6.4497	TRIPLES	1.083	3.7023	8	0.8829
LEFT FIELD	Y=0.00493**RUNS**-5.7021	RUNS	1.005	17.4021	8	0.0262
CENTER FIELD	Y=43.0637**OPS**-36.4753	OPS	>999.999	10.1656	8	0.2356
RIGHT FIELD	Y=0.0555**INTENTIONALWALKS**-9.4257	INTENTIONAL_WALKS	1.057	0.5249	8	0.9998

(Continued)

Time Period Analysis

Pitchers

Models were found that analyze pitchers by the time eras they played the majority of their careers in. They were characterized by many different statistics, which is shown in the analysis below.

Time period one. The model for selected variables filled was chosen as the best fit model because it has better classification and almost the same r-squared value as the original. The r-squared value is strong! 56.36% of variation is explained by the variables indicated in the model. All p-values are less than 0.05. Therefore, not all beta weights are equal to zero. The residual test for goodness of fit looks good. ERA plus was removed because it did not pass the Wald criteria. Our variable SO (strike outs) has a low p-value indicating that it is an important predictor of Hall of Fame admission. For every strike out, the odds of making it into the Hall of Fame increase by a factor of 1.006. Hosmer-Lemeshow also supports the goodness of fit of the model to the data. The correct classification percent is 93.75%.

Time Period Two. The selected variables (original) model was chosen. There is a strong r-squared value! 52.68% of variation is explained by the variables in this model. All p-values are less than 0.05. Therefore, not all beta weights are equal to zero. Therefore, regression continues. The residual test for goodness of fit looks good! Since 0.9932 >0.05, it is concluded that the model fits the data pretty well. Number of walks (variable name "Walks") did not meet criteria to make it into the model and therefore was kicked out. All p-values are less than 0.05 indicating that the variables selected by the program are significant indicators in predicting Hall of Fame for pitchers in time frame two. For every sacrifice hit, the odds of making it into the Hall of Fame increase by a factor of 1.054, all other factors being equal. For every strike out, the odds of

making it into the Hall of Fame increase by a factor of 1.007, all other factors being equal. Hosmer-Lemeshow also supports the goodness of fit of the model to the data. The correct classification percent is 85.71%.

Time Period Three. Selected variables (original) model was selected as the best model found. A strong r-squared value was calculated. 45.76% of variation is explained by the variables in this model. All p-values are less than 0.05. Therefore, it is concluded that not all beta weights are equal to zero.

The residual test for goodness of fit looks good. A p-value of 0.4469 supports this decision, since 0.4469 > 0.05. Number of Years played did not make the cut into the model with a p-value greater than 0.05. All other p-values for remaining variables look good. For every home run, the odds of making it into the Hall of Fame increase by a factor of 1.136, all other factors being equal. For every field percent, the odds of making it into the Hall of Fame increase by a factor of >999, all other factors being equal. For every strike out, the odds of making it into the Hall of Fame increase by a factor of 1.006, all other factors being equal. For every point on ERA Plus, the odds of making it into the Hall of Fame increase by a factor of 1.247, all other factors being equal. The Hosmer-Lemeshow goodness of fit test also supports this model. The correct classification percent is 92.57%.

Time Period Four. The selected variables (original) model was selected as the model of best fit. The r-squared value is moderate. 39.77% of variation is explained the variables chosen in this model. Although the Wald statistic is greater than 0.05, the other tests (Likelihood Ratio and Score) had small p-values; therefore, the analysis is continued. The residual test says ‘yes’ to goodness -of- fit. Number of Saves (variable: SV) no longer made the cut to make it into the

model, so it was removed. Besides the intercept, all variables in the model were important in predicting hall of fame entry. For every game finished, the odds of making it into the Hall of Fame increase by a factor of 1.022, all other factors being equal. For every complete game pitched, the odds of making it into the Hall of Fame increase by a factor of 1.096, all other factors being equal. For every WHIP, the odds of making it into the Hall of Fame decrease by a factor of <0.001, all other factors being equal. Hosmer-Lemeshow also supports our goodness of fit for the model to the data. The correct classification percent is 97.92%.

Catchers

For catchers, the only time period that a model was found was for time period three. In this regression, the validity of the model fit was deemed questionable because of a "complete separation of data points". Taking this into consideration and looking at the model fit statistics and r-squared values, the model is said to be okay and can be considered as a model for the data. The r-squared value is moderate. 39.57% of variation is explained by the variables in the model selected. All p-values meet the 0.05 criteria that the beta weights are not all zero. The residual test for goodness of fit supports the model fit to the data. Lots of variables were stepped in, but many were found to not meet significant criteria. In the end, only On-base plus slugging percent was found to be an important predictor of Hall of Fame entry. For every OPS, the odds of making it into the Hall of Fame increase by a factor of >999. Hosmer-Lemeshow supports the model fit to the data. The correct classification percent was 95.12%.

First Basemen

No models were found for time periods 1, 2, and 4. Time period 3 didn't have a model until the missing values were filled in with the series mean. This was a red flag, however, after careful

consideration of model fit statistics and classification, it is the model that was chosen. A strong r-squared value was found for the model. 55.79% of variation was explained by the variables found by this model. All tests had a p-value within the 0.05 limit. Therefore it is concluded that not all of the beta weights are equal to zero. The residual test for goodness of fit supports the model fit for the data. Triples was removed as a predictor because it no longer met the criteria to stay in the model (p-value <0.05). Slugging percent is the only variable left in the model that meets the significance test. For every slugging percent, the odds of making it into the Hall of Fame increase by a factor of >999. Hosmer-Lemeshow goodness of fit test also supports the model fit to the data. The correct classification percent is 94.44%.

Second Basemen

Only time period 3 had a model appear when the analyses were run. Therefore it is the only one presented. The model selected was select variables filled. The filled model was chosen because it had a comparable r-squared value and a better classification. The r-squared value is pretty strong. 47.11% of variation is explained by the variables in the model. All tests had p-values within 0.05 or less. Therefore, it is concluded that not all beta weights are equal to zero. The residual test for goodness of fit supports the model for the data. Fielding percent was no longer found to be a significant predictor in the model. Slugging percent was the only predictor found to be important by the analysis. For every slugging percent, the odds of making it into the Hall of Fame increase by a factor of >999. Hosmer-Lemeshow also supports the model fit to the data. The correct classification percent was 93.55%.

Third Basemen

No model was found for any of the four time periods for third basemen.

Short Stop

No models were found for time periods 1, 2, and 4. Only models were found in time period three once missing data was filled in with the series mean. After careful analysis, the model chosen was select variables filled. The r-squared value is moderate. 35.68% of variation is explained by the variables in the model. All p-values are less than 0.05 for all tests. Therefore, it is concluded that not all beta weights are equal to zero. The residual test for goodness of fit supports the model. Field percent no longer meets the criteria to be kept in the model. Triples are the only factor left in the model. For every triple, the odds of making it into the Hall of Fame increase by a factor of 1.083. Hosmer-Lemeshow goodness of fit test also supports the model. The correct classification percent is 87.18%.

Left Fielder

No models were found for time periods 1, 2, and 4. For time period 3, the selected model was select variables (original). The r-squared value is pretty strong. 41.51% of variation is explained by the variables in the model. All p-values are under the 0.05 limit. Therefore, it is concluded that not all beta weights are equal to zero. The residual test does not support the model fit to the data. However, other tests were run to check for model fit. Runs was the only variable deemed important by the stepwise regression. For every run, the odds of making it into the Hall of Fame increase by a factor of 1.005. The Hosmer-Lemeshow test does not support the model. However, the correct classification percent is 91.89%.

Center Fielder

There were no models found in time periods 1, 2, and 4. The model chosen for time period 3 was the selected variables (original). Validity is questionable because of complete separation of data points. However, after looking at the output, the model seems to fit the data pretty well. The r-squared value is strong. 53.21% of variation is explained by the variables in the model. All p-values are within the 0.05 limits. Therefore, it is concluded that not all beta weights are equal to zero. The residual test for goodness of fit supports the model. Pick offs were found to no longer be an important predictor for hall of fame entry for center fielders. On base plus slugging percent was found to be the only significant predictor in hall of fame entry. For every OPS percent, the odds of making it into the Hall of Fame increase by a factor of >999. Hosmer-Lemeshow also supports the model's goodness of fit. The correct classification percent was 86.49%.

Right Fielder

No models for any time period were found except for time period 4. Select variables are picked as the model. Validity is questionable; however the classification is quite strong, so the model is used. The r squared value is low to moderate. 28.44% of variation is explained by the variables in the model. All p-values are less than 0.05. Therefore, it is concluded that not all beta weights are equal to zero. The residual test supports the model. "Stolen bases" was removed from the model because it no longer fit the 0.05 criteria. "Intentional walks" is the only variable was found to be in the model. For every intentional walk, the odds of making it into the Hall of Fame increase by a factor of 1.057. Hosmer-Lemeshow also supports the goodness of fit for the model. The correct classification percent is 88.10%.

Time frame analysis did not turn out the way that it was planned. In a perfect world, a model for each position and every time frame would be available. However, this is not the case. Data set size could be a factor in this. The goal was to compare the different time periods within positions to see similarities and differences. However, this cannot be done since there are time periods missing models. The only position that can be compared is pitchers. Although not the only predictors, strikeouts appeared in both models for time periods 1, 2, and 3. As expected, the model for time period 4 used modern statistics, such as WHIP. For time period 3, first and second basemen both had slugging percent as a predictor. These were the only similarities seen between the different positions.

In summary, models were found to predict hall of fame entry by position using overall data. As anticipated, many models found used RUNS and RBIS as a predictor. Most commonly thought of in hall of fame entry is batting average; however as shown in the model analysis, batting average was not found to be a contributor to hall of fame entry at **any** position. Time period analysis, as previously mentioned, did not turn out as planned. The goal was to find a model for each position during each time period. This was unsuccessful. Instead, for every position, one model was found within one time period, with the exception on third base (which had no models) and pitchers (which had a model for every time period).

CHAPTER V

PREDICTIONS

Players from the current Washington Nationals and Detroit Tigers teams were compiled into a new data set. The data set consists of the starters for the game on September 4, 2013. The regression equations chosen for each position will be used, and the results will show the probability that a person at a given position will make it into the hall of fame. The probabilities were computed in MS Excel 2010.

Table 5.1

Predictions for the starting line-ups for the Tigers and Nationals for September 4, 2013

Name	Position	Probability
J. Zimmermann	1	0.000437129
R. Porcello	1	6.33891E-06
W. Ramos	2	0.001354727
A. Avila	2	0.002428234
A. LaRoche	3	2.609E-06
P. Fielder	3	2.67626E-05
S. Lombardozzi	4	6.01205E-09
O. Infante	4	2.02508E-06
R. Zimmerman	5	0.117377744
M. Cabrera	5	0.962013586
I. Desmond	6	2.90536E-05
J. Iglesias	6	7.26258E-07
C. Brown	7	0.002206916
D. Kelly	7	0.003684974
D. Span	8	1.31873E-05
A. Jackson	8	5.55011E-05
J. Werth	9	2.23981E-05
A. Dirks	9	1.56386E-07

Out of the 18 starters listed above, only one was predicted to make it into the hall of fame. That person is third baseman Miguel Cabrera from the Detroit Tigers. He would be classified as a "shoe in" into the hall of fame.

Table 5.2

Predictions for the top 25 players in MLB

RANK	PLAYER	TEAM	POSITION	PROBABILITY
3	**Clayton Kershaw**	LAD	1	0.011745839
10	**Max Scherzer**	DET	1	0.000577513
14	**Jose Fernandez**	MIA	1	0.097996981
15	**Matt Harvey**	NYM	1	0.014477696
18	**Yu Darvish**	TEX	1	0.002249928
20	**Chris Sale**	CHW	1	0.004863031
22	**Greg Holland**	KC	1	0.029783303
24	**Koji Uehara**	BOS	1	0.093027482
4	**Chris Davis**	BAL	3	1.77033E-07
7	**Joey Votto**	CIN	3	6.07156E-07
8	**Paul Goldschmidt**	ARI	3	5.58369E-08
17	**Freddie Freeman**	ATL	3	6.2839E-08
19	**David Ortiz**	BOS	3	0.29596018
6	**Matt Carpenter**	STL	4	1.94131E-08
12	**Robinson Cano**	NYY	4	0.290128508
16	**Jason Kipnis**	CLE	4	3.1642E-08
23	**Dustin Pedroia**	BOS	4	0.00021108
1	**Miguel Cabrera**	DET	5	0.962013586
11	**Edwin Encarnacion**	TOR	5	0.012115263
13	**Josh Donaldson**	OAK	5	0.002879084
21	**Adrian Beltre**	TEX	5	0.895612696
25	**Kyle Seager**	SEA	5	0.010726007
2	**Mike Trout**	LAA	8	0.709006652
5	**Andrew McCutchen**	PIT	8	0.060114957
9	**Shin-Soo Choo**	CIN	8	0.021892459

The table above shows the top 25 players in MLB and the probabilities that they are going to make it into the Hall of Fame. Of these players, only 3 were given probabilities over 50%: Miguel Cabrera, Adrian Beltre, and Mike Trout. What this analysis is saying is that if the players keep up their current numbers, they are highly probable to make it into the hall of fame. As for the other players, there are other factors that could possibly be influencing the results. One reason could be that the numbers aren't there yet. Many of the players listed are within the first 5 years of their careers. They haven't had the time to rack up their numbers.

Time Period Analysis

The idea was to have players from the different eras that are in the hall of fame be chosen and each will be put through a model to see if being evaluated for the hall of fame in a different time would have affected the outcome of being admitted into the baseball hall of fame. However, this cannot happen except for pitchers. For the positions that had one equation (total), players that are from the other time periods will be run through the model to see if they would make it or not. One player from each time period and each position was chosen to be put to the test. Results can be found on the following few pages.

Table 5.3

Predictions for Pitchers to make it into the Hall of Fame for different Time Periods

Pitcher				
Time Period being Analyzed	**Name**	**Position**	**Time Period Player Played**	**Probability**
1	**Burleigh Grimes**	**1**	**2**	0.80100062
	Sandy Koufax	**1**	**3**	0.99828988
	Goose Gossage	**1**	**4**	0.23196657
2	**Jack Chesbro**	**1**	**1**	0.08314601
	Sandy Koufax	**1**	**3**	0.99883632
	Goose Gossage	**1**	**4**	0.00759305
3	**Jack Chesbro**	**1**	**1**	8.9244E-05
	Burleigh Grimes	**1**	**2**	0.00080811
	Goose Gossage	**1**	**4**	1.2162E-05
4	**Jack Chesbro**	**1**	**1**	0.99998773
	Burleigh Grimes	**1**	**2**	0.99986561
	Sandy Koufax	**1**	**3**	0.83273404

The four players chosen to run through the time period analysis are all well-known players. Sandy Koufax is predicted to make it into the hall of fame regardless of time era. Burleigh Grimes was predicted to make it in time period 1 and 4. Goose Gossage was predicted to **not** make it into the hall of fame in any time period. Jack Chesbro was predicted to make it into the hall of fame in time period 4.

Table 5.4

Predictions for Catchers to Make it into the Hall of Fame for Different Time Periods

Catcher				
Time Period being Analyzed	**Name**	**Position**	**Time Period Player Played**	**Probability**
3	**Buck Ewing**	**2**	**1**	0.44045482
	Ray Schalk	**2**	**2**	0.00162524
	Gary Carter	**2**	**4**	0.16367978

For catchers, only a model for time period 3 was found. The highest probability of making it in was for Buck Ewing at 44%. However by the 50% criteria established, he is said to not make it in. Ray Shalk and Gary Carter didn't come close.

Table 5.5

Predictions for First Basemen to Make it into the Hall of Fame for Different Time Periods

First Baseman				
Time Period being Analyzed	**Name**	**Position**	**Time Period Player Played**	**Probability**
3	**Jake Beckley**	**3**	**1**	0.001310706
	George Kelly	**3**	**2**	0.005523618
	Eddie Murray	**3**	**4**	0.046126581

For first basemen, only a model for time period 3 was found. The highest probability of making it in was for Eddie Murray at 4.6%. This is a **terrible** percentage. Not only that, but Jake Beckley and George Kelly were less than that!

Table 5.6

Predictions for Second Basemen to Make it into the Hall of Fame for Different Time Periods

Second Baseman				
Time Period being Analyzed	**Name**	**Position**	**Time Period Player Played**	**Probability**
3	**Bid McPhee**	**4**	**1**	0.100719289
	Eddie Collins	**4**	**2**	0.707860022
	Rod Carew	**4**	**4**	0.707860022

For second basemen, only a model for time period 3 was found. The highest probability of making it in was for Eddie Collins and Rod Carew both at 70.79%. Bid McPhee didn't quite make the cut with only a 10% probability.

There were no models found for any time period for third basemen. Why this is? That is for another analysis.

Table 5.7

Predictions for Short Stops to Make it into the Hall of Fame for Different Time Periods

Short Stop				
Time Period being Analyzed	**Name**	**Position**	**Time Period Player Played**	**Probability**
3	**Honus Wagner**	**6**	**1**	0.9999988
	Rabbit Maranville	**6**	**2**	0.99952759
	Cal Ripken, Jr.	**6**	**4**	0.05007315

For short stops, only a model for time period 3 was found. The highest probability of making it in was for Honus Wagner at almost 100%! A close second was Rabbit Maranville with 99.95% probability based on this model. It's interesting, though, that Cal Ripken Jr. only had a 5% probability of making it into the hall of fame in time period three. Who doesn't know of Cal Ripken Jr., and anyone who does *knows* that he was an amazing athlete (he was inducted into the hall of fame in time period 4…)

Table 5.8

Predictions for Left fielders to Make it into the Hall of Fame for Different Time Periods

Left Fielder				
Time Period being Analyzed	**Name**	**Position**	**Time Period Player Played**	**Probability**
3	**Joe Kelley**	**7**	**1**	0.78641168
	Goose Goslin	**7**	**2**	0.833285007
	Jim Rice	**7**	**4**	0.611938986

For left fielders, only a model for time period 3 was found. The highest probability of making it in was for Goose Goslin at 83.33%. Joe Kelley and Jim Rice were also given good probabilities, 78.64% and 61.19% respectively.

Table 5.9

Predictions for Center Fielders to Make it into the Hall of Fame for Different Time Periods

Center Fielder				
Time Period being Analyzed	**Name**	**Position**	**Time Period Player Played**	**Probability**
3	**Hugh Duffy**	**8**	**1**	0.393891601
	Hack Wilson	**8**	**2**	0.982094472
	Kirby Puckett	**8**	**4**	0.393891601

For center fielders, only a model for time period 3 was found. The highest probability of making it in was for Hack Wilson at 98.21%. Hugh Duffy and Kirby Puckett did not make the cut, both with 39.39% probability of making it into the hall of fame in time period three.

Table 5.10

Predictions for Right Fielders to Make it into the Hall of Fame for Different Time Periods

Right Fielder				
Time Period being Analyzed	**Name**	**Position**	**Time Period Player Played**	**Probability**
4	**Tommy McCarthy**	**9**	**1**	n/a
	Babe Ruth	**9**	**2**	n/a
	Hank Aaron	**9**	**3**	0.998926546

For right fielders, only a model for time period 4 was found. The highest probability of making it in was for Hank Aaron at 99.89%. For right fielders, intentional walks were not tracked in time periods one and two; therefore probabilities were unable to be computed for Tommy McCarthy and Babe Ruth.

The goal for the time frame analysis was to show how different players were evaluated based on different statistics to make it into the hall of fame. However this goal was unable to be achieved due to the lack of models found.

CHAPTER VI

FURTHER STUDIES, REMARKS, AND CONCLUSION

The study completed in this thesis was solely based on statistics. If statistics were the only item taken into consideration when evaluating whether a player gets inducted into the hall of fame, then the models produced can predict this. However, induction is not based completely on statistics.

An important remark is that although this thesis research has been carefully observed and monitored, it does not mean that there are not data problems, either from the website itself (where the data was obtained) or from data entry. Baseball-reference.com does not guarantee complete accuracy of data and cannot be held responsible for discrepancies. Data entry and compilation may also be flawed, which would make the models determined somewhat inaccurate.

For baseball enthusiasts, baseball has nine different positions. However, pitchers can be split up into categories: starting pitchers ("in the rotation"), relief pitchers, and closers. For the purposes of this thesis, pitchers were grouped into one category and not by their playing time in the game. If these players were re-grouped based upon their role in the game, models would likely be adjusted.

Another important remark is that although many different models were found using this data, they are not the only models that exist to predict the Hall of Fame. If data was compiled differently or if different statistics were chosen, the models would be

different and, therefore, the probabilities would be different. A player could have a probability of 80 percent with a model constructed using the methods described in this thesis, and yet not make the cut for a model constructed through another source.

According to the baseball hall of fame website, players are voted "based upon the player's record, playing ability, integrity, sportsmanship, character, and contributions to the team(s) which the player played" (baseballhall.com). According to Graham Womack, among the long list of players that should be in the hall but are not is Shoeless Joe Jackson. His numbers were there, however he compromised the integrity of the game when he accepted money from gamblers to throw the world series of 1919. Although not as big of an issue now, another issue has recently emerged that is in high debate for ballot entry now: Steroids and Human Growth Hormone (HGH). Many players in the nineties and early 2000's used steroids and/or HGH and were up for ballot entry in the recent election years. However because of this trend, players were not voted into the hall. Is taking drugs to make a swing more powerful or a runner faster any better than what Joe Jackson did? This is a question that the hall of fame needs to answer soon because this problem is not going to go away any time soon.

As for further studies, it would be interesting to look at players who were known to have taken steroids/HGH and see how their game changed by looking at their numbers over their careers. Did Barry Bonds show any spikes in his batting and fielding when he was consistently taking performance enhancing drugs? Also looking at how players

game's changed when they were rotated into different positions would be interesting. Did Michael Young's game change when he was pulled from second, moved to short and then to third? Comparing fielding statistics, and even batting statistics, would be interesting to see. Maybe one area of the game negatively affects the other. The only way to know is to run an analysis and see the differences in the statistics.

The purpose of this thesis was to analyze players that were nominated and inducted into the Baseball Hall of Fame. Using logistic regression, models were obtained and current Major League Baseball players were evaluated. These players were then given a probability that they would be inducted into the hall of fame. This information was trying to further sabermetric studies. Knowing and understanding the methods provided in this thesis, it is believed that this goal was achieved. For non-sabermetricians, this thesis satisfied the curiosity of the average baseball fan.

REFERENCES

ESPN. (2013). Retrieved September 7, 2013, from MLB Player Ratings: http://espn.go.com/mlb/playerratings

Baranger, W. (2007). *Gamblers and Miscreants*. Retrieved May 14, 2013, from What the Hall: http://what-the-hall.info/index.shtml?gamblers

BBWAA. (n.d.). Retrieved May 13, 2013, from National Baseball Hall of Fame and Museum: http://baseballhall.org/hall-famers/rules-election/bbwaa

BBWAA. (n.d.). Retrieved September 6, 2013, from National Baseball Hall of Fame and Museum: http://baseballhall.org/hall-famers/rules-election/bbwaa

Birnbaum, P. (n.d.). *A Guide to Sabermetric Research*. Retrieved May 14, 2013, from Society for American Baseball Research: http://sabr.org/sabermetrics

Brian Mills, S. S. (2011). Using Tree Ensembles to Analyze National Baseball Hall of Fame Voting Patterns: An Application to Discrimination in BBWAA. *Journal of Quantitative Analysis in Sports*.

Grajkowske, A. (2008). *Using Binary Logistic Regression to Predict Baseball Hall of Fame Admittance.* Rapid City.

Hall of Fame Voting Procedures. (n.d.). Retrieved May 14, 2013, from Baseball Reference: http://www.baseball-reference.com/about/hof_voting.shtml

Hosmer, D., & Lemeshow, S. (2000). *Applied Logistic Regression.* Danvers: John Wiley and Sons Inc.

James, B. (1988). *A Bill James Primer.* Retrieved May 15, 2013, from The Baseball Archive: http://web.archive.org/web/20050621004022/http://www.baseball1.com/bb-data/bbd-bj1.html

Johnson, D. E. (2013, July 16). Retrieved November 3, 2013, from DEJ on Edge: http://dejonedge.blogspot.com/

Johnson, T. (n.d.). *The History of the Baseball.* Retrieved August 7, 2013, from Fog Dog Sports: http://fogdogsports.com/fog-the-history-of-the-basbeall--bg-527303.html

Kleinbaum, D., Klein, M., & Pryor, E. R. (2002). *Logistic Regressiom: A Self Learning Text.* New York: Springer.

Pallant, J. (2006). *SPSS Survival Manual: 2nd Edition.* New York: Open University Press.

Richard Picard, D. C. (1984). Cross-Validation of Regression Models. *Journal of the American Statistial Association*, 575-583.

Womack, G. (2011, December 11). *The 50 Best Baseball Players not in the Hall of Fame*. Retrieved September 6, 2013, from Baseball: Past and Present: http://baseballpastandpresent.com/2011/12/11/50-baseball-players-hall-fame-version-2-0/

APPENDIX A

ADDITIONAL REGRESSION INFORMATION

ADDITIONAL REGRESSION INFORMATION

PITCHERS										
TYPE		**WALD:GLOBAL NULL BETA=0**			**RESIDUAL**			**HOSMER-LEMESHOW**		
	R^2	**X^2**	**DF**	**P-VALUE**	**X^2**	**DF**	**P-VALUE**	**X^2**	**DF**	**P-VALUE**
ORIGINAL	0.4786	52.4991	8	0.0001	47.2172	48	0.5048	153.7804	8	0.0001
ORIGINAL:ENTER METHOD	0.4241	59.7005	5	0.0001	0.7137	2	0.6999	54.1617	8	0.0001
ORIGINAL: PITCHING STATS ONLY	0.3711	31.4312	3	0.0001	32.0679	27	0.2296	41.1995	8	0.0001
ORIGINAL: PITCHING STATS ONLY:FILLED	0.4899	56.9508	8	0.0001	20.4684	21	0.4918	628.0456	8	0.0001
ADDED	0.3988	32.8269	4	0.0001	51.5676	56	0.6432	202.9661	8	0.0001
ADDED AND FILLED	0.2345	62.9704	1	0.0001	296.4307	56	0.6432	32.4287	8	0.0001
FILLED:ALL VARIABLES	0.2345	62.9704	1	0.0001	311.0089	69	0.0001	32.4287	8	0.0001

PITCHERS		
TYPE	**MODEL**	**CLASSIFICATION PERCENT**
ORIGINAL	Y= 0.00274**ATBATS**-0.2171**TRIPLES**+0.0642**WINS**-0.0151**GS_1**+0.0159**CG_1**+0.0121**SAVES**-0.0216**WP_1**-2.4147**H_9**+8.9298	93.68421053
ORIGINAL:ENTER METHOD	Y= -0.1747**TRIPLES**+0.0620**WINS-**0.0179**GS_1**+0.0189**CG_1**+0.00856**SAVES-**9.5809	92.36842105
ORIGINAL: PITCHING STATS ONLY	Y=0.0417**CG_1+**0.0156**SAVES**-2.5264**H_9**+12.6874	90
ORIGINAL: PITCHING STATS ONLY:FILLED	Y=0.0463**WINS**-0.0118**GS_1**+0.0314**CG_1+**0.0151**SAVES**-0.00584**ER**+0.00545**BB**-0.0591**ERA**-21.4271**WHIP**+23.5898	87.10526316
ADDED	Y=0.0423**CG**+17.6976**WLPERCENT**+0.0223**SAVES**-18.3097**WHIP**+4.1270	91.31578947
ADDED AND FILLED	Y=0.0942**SHO**-3.9200	87.89473684
FILLED:ALL VARIABLES	Y=0.0942**SHO**-3.9200	87.89473684

CATCHERS										
TYPE		**WALD:GLOBAL NULL BETA=0**			**RESIDUAL**			**HOSMER-LEMESHOW**		
	R^2	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**
ORIGINAL	0.2521	8.9577	1	0.0002	31.2552	30	0.4029	6.6717	9	0.6713
ADDED	0.2521	8.9577	1	0.0028	32.2084	32	0.4564	6.6717	9	0.6713
ADDED AND FILLED					85.0322	34	0.0001			
FILLED:ALL VARIABLES					88.175	50	0.0007			

CATCHERS		
TYPE	**MODEL**	**CLASSIFICATION PERCENT**
ORIGINAL	Y=0.00696**RUNS**-7.1944	86.86868687
ADDED	Y=0.00696**RUNS**-7.1944	86.86868687
ADDED AND FILLED	Y=-1.6463	
FILLED:ALL VARIABLES	Y=-1.6463	

FIRST BASEMAN										
TYPE		**WALD:GLOBAL NULL BETA=0**			**RESIDUAL**			**HOSMER-LEMESHOW**		
	R^2	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**
ORIGINAL	0.4352	5.5019	3	0.1385	19.6454	27	0.8453	1.5557	7	0.9803
ADDED	0.3531	8.2547	2	0.0161	26.4326	29	0.6023	0.6345	8	0.9997
ADDED AND FILLED	0.2603	16.536	1	0.0001	63.6525	31	0.0005	9.7326	6	0.1364
FILLED:ALL VARIABLES	0.2603	16.536	1	0.0001	77.9956	40	0.0003	9.7326	6	0.1364

FIRST BASEMEN		
TYPE	**MODEL**	**CLASSIFICATION PERCENT**
ORIGINAL	Y=-0.00379**PLATEAPPEARANCES**+0.0975**TRIPLES**+0.0334**RBIS**-18.9221	90.0990099
ADDED	Y=0.0126**RBIS**-0.0291**GDP**-13.7434	89.10891089
ADDED AND FILLED	Y=70.0275**BATTINGAVERAGE**-22.0998	88.11881188
FILLED:ALL VARIABLES	Y=70.0275**BATTINGAVERAGE**-22.0998	88.11881188

SECOND BASEMAN										
TYPE		WALD:GLOBAL NULL BETA=0			RESIDUAL			HOSMER-LEMESHOW		
	R^2	χ^2	DF	P-VALUE	χ^2	DF	P-VALUE	χ^2	DF	P-VALUE
ORIGINAL	0.4611	11.8744	2	0.0026	21.5578	27	0.7595	5.3824	8	0.716
ADDED	0.5071	5.9157	3	0.1158	13.8776	27	0.9824	1.581	7	0.9794
ADDED AND FILLED	0.366	18.3831	2	0.0001	34.4282	29	0.2239	6.0699	8	0.6394
FILLED:ALL VARIABLES	0.366	18.3831	2	0.0001	42.9957	39	0.304	6.0699	8	0.6394

SECOND BASEMEN		
TYPE	MODEL	CLASSIFICATION PERCENT
ORIGINAL	Y=-0.00661**STRIKEOUTS**+0.00446**TOTALBASES**-9.2366	89.15662651
ADDED	Y=0.0236**RBIS**+0.0347**GDP**-0.00924**STRIKEOUTS**-19.5313	93.97590361
ADDED AND FILLED	Y=0.0283**TRIPLES**+28.2697**SLUGGINGPERCENT**-14.8745	83.13253012
FILLED:ALL VARIABLES	Y=28.2697**SLUGGINGPERCENT**+0.0283**TRIPLES**-14.8745	83.13253012

THIRD BASEMAN										
TYPE		**WALD:GLOBAL NULL BETA=0**			**RESIDUAL**			**HOSMER-LEMESHOW**		
	R^2	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**
ORIGINAL	0.3922	11.6172	2	0.0038	25.7265	25	0.4223	2.934	8	0.9384
ADDED	0.3394	11.9732	1	0.0005	37.7926	28	0.1024	9.1727	8	0.3279
ADDED AND FILLED	0.3087	19.2028	1	0.0001	55.1814	30	0.0034	4.2165	5	0.5187
FILLED:ALL VARIABLES	0.3087	19.2028	1	0.0001	62.4321	40	0.0131	4.2165	3	0.5187

THIRD BASEMEN		
TYPE	**MODEL**	**CLASSIFICATION PERCENT**
ORIGINAL	Y=-1.0770**YEARS**+0.0101**HITS**-4.8687	85.39325843
ADDED	Y=0.00702**RUNS**-8.7715	80.8988764
ADDED AND FILLED	Y=85.0314**BATTINGAVERAGE**-25.4313	89.88764045
FILLED:ALL VARIABLES	Y=85.0314**BATTINGAVERAGE**-25.4313	89.88764045

SHORT STOP										
TYPE		**WALD:GLOBAL NULL BETA=0**			**RESIDUAL**			**HOSMER-LEMESHOW**		
	R^2	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**
ORIGINAL	0.3503	17.7123	2	0.0001	31.4467	24	0.1413	4.4433	8	0.8151
ADDED	0.5053	12.0104	2	0.0025	24.8158	27	0.5848	3.6533	8	0.887
ADDED AND FILLED	0.2377	14.7345	1	0.0001	74.5704	29	0.0001	18.7126	7	0.0091
FILLED:ALL VARIABLES	0.2377	14.7345	1	0.0001	79.0841	4	0.0002	18.7126	7	0.0091

SHORT STOP		
TYPE	**MODEL**	**CLASSIFICATION PERCENT**
ORIGINAL	Y=0.00457**RIBS**+27.7546**ONBASEPERCENT**-14.0695	82.35294118
ADDED	Y=0.00905**RBIS**+0.00202**PO**-14.5948	87.05882353
ADDED AND FILLED	Y=0.00384**RBIS**-3.8240	76.47058824
FILLED:ALL VARIABLES	Y=0.00384**RBIS**-3.8240	76.47058824

LEFT FIELDER										
TYPE		**WALD:GLOBAL NULL BETA=0**			**RESIDUAL**			**HOSMER-LEMESHOW**		
	R^2	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**
ORIGINAL	0.2459	13.8498	1	0.0002	34.1667	30	0.2742	15.2813	8	0.0539
ORIGINAL: ENTER	0.3356	19.3368	1	0.0001				23.4475	8	0.0028
ADDED	0.2459	13.8498	1	0.0002	34.1667	30	0.2742	15.2813	8	0.0539
ADDED AND FILLED	0.4065	21.6913	3	0.0001	32.8074	27	0.2035	27.2375	8	0.0006
FILLED:ALL VARIABLES	0.3973	19.7357	2	0.0001	42.7388	39	0.3137	17.2356	8	0.0277

LEFT FIELDER		
TYPE	**MODEL**	**CLASSIFICATION PERCENT**
ORIGINAL	Y=0.00378**RUNS**-5.5018	86.95652
ORIGINAL: ENTER	Y=0.00455**RUNS**-6.1367	89.13043
ADDED	Y=0.00378**RUNS**-5.5018	86.95652
ADDED AND FILLED	Y=0.00438**RUNS**+0.02**TRIPLES**-0.00322**STOLENBASES**-7.0018	90.21739
FILLED:ALL VARIABLES	Y=0.00619**RUNS**-0.0112**INTENTIONALWALKS**-7.3620	91.30435

CENTER FIELDER										
TYPE		WALD:GLOBAL NULL BETA=0			RESIDUAL			HOSMER-LEMESHOW		
	R^2	χ^2	DF	P-VALUE	χ^2	DF	P-VALUE	χ^2	DF	P-VALUE
ORIGINAL	0.5227	7.9251	2	0.019	23.5638	25	0.5447	1.4151	8	0.994
ADDED	0.5227	7.9251	2	0.019	23.8113	27	0.6408	1.4151	8	0.994
ADDED AND FILLED					77.1742	30	0.0001			
FILLED:ALL VARIABLES					79.1635	41	0.0003			

CENTER FIELD		
TYPE	MODEL	CLASSIFICATION PERCENT
ORIGINAL	Y=58.8519**OPS**+0.0836**E**-55.6474	87.09677419
ADDED	Y=58.8519**OPS**+0.0836**E**-55.6474	87.09677419
ADDED AND FILLED	Y=-1.1130	
FILLED:ALL VARIABLES	Y=-1.1130	

RIGHT FIELDER										
TYPE		WALD:GLOBAL NULL BETA=0			RESIDUAL			HOSMER-LEMESHOW		
	R^2	χ^2	DF	P-VALUE	χ^2	DF	P-VALUE	χ^2	DF	P-VALUE
ORIGINAL	0.539	9.4574	2	0.0088	67.1389	28	0.0001	9.1616	7	0.2413
ADDED	0.3563	12.4046	1	0.0004	33.2161	29	0.2691	6.1849	8	0.6265
ADDED AND FILLED	0.5828	24.4519	28	0.6575	10.9392	3	0.0121	0.3995	8	0.9999
FILLED:ALL VARIABLES	0.5828	10.9392	3	0.0121	24.9705	38	0.9486	0.3995	8	0.9999

RIGHT FIELDER		
TYPE	MODEL	CLASSIFICATION PERCENT
ORIGINAL	Y=0.00893**RUNS**+0.0632**TRIPLES**-17.3814	94.68085106
ADDED	Y=0.0634**TRIPLES**-6.1316	86.17021277
ADDED AND FILLED	Y=0.00811**RUNS**+0.0657**TRIPLES**+0.0230**E**-17.6574	94.68085106
FILLED:ALL VARIABLES	Y=0.00811**RUNS**+0.0657**TRIPLES**+0.0230**E**-17.6574	94.68085106

PITCHERS											
TIME PERIOD 1			WALD:GLOBAL NULL BETA=0			RESIDUAL			HOSMER-LEMESHOW		
		R^2	χ^2	DF	P-VALUE	χ^2	DF	P-VALUE	χ^2	DF	P-VALUE
	SELECT VARIABLES	0.5789	5.3631	1	0.0206	11.8264	25	0.9879	1.1297	6	0.9802
	SELECT VARIABLES:FILLED	0.5636	7.6414	1	0.0057	19.4421	30	0.9303	2.3876	9	0.9838
	ALL VARIABLES:FILLED	0.5636	7.6414	1	0.0057	19.4421	30	0.9303	2.3876	9	0.9838
TIME PERIOD 2	SELECT VARIABLES	0.5268	10.2577	2	0.0059	26.4724	47	0.9932	4.6006	8	0.7993
	SELECT VARIABLES:FILLED	0.4454	13.7566	2	0.001	37.5976	53	0.9458	3.3147	7	0.8544
	ALL VARIABLES:FILLED	0.4454	13.7566	2	0.001	37.5976	53	0.9458	3.3147	7	0.8544
TIME PERIOD 3	SELECT VARIABLES	0.4576	12.3296	4	0.0151	59.7871	59	0.4469	4.378	8	0.8215
	SELECT VARIABLES:FILLED	0.4704	19.9977	4	0.0005	156.1194	55	0.0001	120.3835	8	0.0001
	ALL VARIABLES:FILLED	0.166	20.1933	1	0.001	211.455	67	0.0001	47.7217	8	0.0001
TIME PERIOD 4	SELECT VARIABLES	0.3977	6.8283	3	0.0776	15.4371	58	1	1.0804	6	0.9824
	SELECT VARIABLES:FILLED	0.3871	0.9521	3	0.8868	15.4542	57	1	0.5229	6	0.9975
	ALL VARIABLES:FILLED	0.3871	6.9521	3	0.0734	15.4657	65	1	0.5229	6	0.9975

PITCHERS		
TIME PERIOD 1	**MODEL**	**CLASSIFICATION PERCENT**
	Y=0.2412**SHO**-8.2353	84.375
	Y=0.00563**SO**-7.12	93.75
	Y=0.00563**SO**-7.12	93.75
TIME PERIOD 2	Y=0.0521**SACRIFICEHITS**+0.00745**SO**-12.9187	85.71428571
	Y=0.0609**SACRIFICEHITS**+3.9676**SO/BB**-9.9068	85.71428571
	Y=0.0609**SACRIFICEHITS**+3.9676**SO/BB**-9.9068	85.71428571
TIME PERIOD 3	Y=0.1272**HOMERUNS**+180.4**FIELDPERCENT**+0.00585**SO**+0.2211**ERAPLUS**-212.2	92.56756757
	Y=-0.6747**YEARS**-38.9047**FIELDPERCENT**+0.0608**WINS**+0.2128**ERAPLUS**+5.5491	87.75510204
	Y=0.0179**WINS**-4.4754	89.11564626
TIME PERIOD 4	Y=0.0214**GAMESFINISHED**+0.092**CG_1**-38.7818**WHIP**+30.9522	97.91666667
	Y=0.0940**COMPLETEGAMES**+0.0220**GAMESFINISHED**-39.3918**WHIP**+31.3432	98.61111111
	Y=0.0940**COMPLETEGAMES**+0.0220**GAMESFINISHED**-39.3918**WHIP**+31.3432	98.61111111

\

CATCHERS											
TIME PERIOD 1			**WALD:GLOBAL NULL BETA=0**			**RESIDUAL**			**HOSMER-LEMESHOW**		
		R^2	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**
	SELECT VARIABLES					12	11	0.3636			
	SELECT VARIABLES:FILLED					15	14	0.3782			
	ALL VARIABLES:FILLED					15	14	0.3782			
TIME PERIOD 2	**SELECT VARIABLES**					16	15	0.3821			
	SELECT VARIABLES:FILLED					18	17	0.3888			
	ALL VARIABLES:FILLED					18	17	0.3888			
TIME PERIOD 3	**SELECT VARIABLES**	0.3957	6.0011	1	0.0143	24.1101	31	0.806	5.0881	8	0.7481
	SELECT VARIABLES:FILLED	0.4627	6.7812	1	0.0092	15.317	32	0.9944	3.1004	6	0.7961
	ALL VARIABLES:FILLED	0.4627	6.7812	1	0.0092	15.3205	39	0.9998	3.1004	6	0.7961
TIME PERIOD 4	**SELECT VARIABLES**					25	24	0.4058			
	SELECT VARIABLES:FILLED					25	24	0.4058			
	ALL VARIABLES:FILLED					25	24	0.4058			

CATCHERS			
TIME PERIOD 1		**MODEL**	**CLASSIFICATION PERCENT**
	SELECT VARIABLES	Y=-1.6094	
	SELECT VARIABLES:FILLED	Y=-1.8718	
	ALL VARIABLES:FILLED	Y=-1.8718	
TIME PERIOD 2	**SELECT VARIABLES**	Y=-2.7081	
	SELECT VARIABLES:FILLED	Y=-2.0794	
	ALL VARIABLES:FILLED	Y=-2.0794	
TIME PERIOD 3	**SELECT VARIABLES**	Y=40.9348**OPS**-33.2737	95.12195122
	SELECT VARIABLES:FILLED	Y=111.8**ONBASEPERCENT**-40.4284	87.80487805
	ALL VARIABLES:FILLED	Y=111.8**ONBASEPERCENT**-40.4284	87.80487805
TIME PERIOD 4	**SELECT VARIABLES**	Y=-1.9924	
	SELECT VARIABLES:FILLED	Y=-1.9924	
	ALL VARIABLES:FILLED	Y=-1.9924	

FIRST BASEMAN											
TIME PERIOD 1			**WALD:GLOBAL NULL BETA=0**			**RESIDUAL**			**HOSMER-LEMESHOW**		
		R^2	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**
	SELECT VARIABLES					10	9	0.3505			
	SELECT VARIABLES:FILLED					11	10	0.3575			
	ALL VARIABLES:FILLED					11	10	0.3575			
TIME PERIOD 2	**SELECT VARIABLES**					14	12	0.3007			
	SELECT VARIABLES:FILLED					15	13	0.3074			
	ALL VARIABLES:FILLED					15	13	0.3074			
TIME PERIOD 3	**SELECT VARIABLES**					29	27	0.3609			
	SELECT VARIABLES:FILLED	0.5579	4.0581	1	0.044	11.5453	30	0.999	7.351	5	0.1958
	ALL VARIABLES:FILLED	0.5579	4.0581	1	0.044	11.5463	33	0.9998	7.351	5	0.1958
TIME PERIOD 4	**SELECT VARIABLES**					38	37	0.4236			
	SELECT VARIABLES:FILLED					38.9993	37	0.38			
	ALL VARIABLES:FILLED					39	38	0.4246			

FIRST BASEMAN			
TIME PERIOD 1		**MODEL**	**CLASSIFICATION PERCENT**
	SELECT VARIABLES	Y=0	
	SELECT VARIABLES:FILLED	Y=-0.1823	
	ALL VARIABLES:FILLED	Y=-0.1823	
TIME PERIOD 2	**SELECT VARIABLES**	Y=-0.9163	
	SELECT VARIABLES:FILLED	Y=-0.6931	
	ALL VARIABLES:FILLED	Y=-0.6931	
TIME PERIOD 3	**SELECT VARIABLES**	Y=-1.5686	
	SELECT VARIABLES:FILLED	Y=90.1684**SLUGGINGPERCENT**-45.9493	94.44444444
	ALL VARIABLES:FILLED	Y=90.1684**SLUGGINGPERCENT**-45.9493	94.44444444
TIME PERIOD 4	**SELECT VARIABLES**	Y=-3.6109	
	SELECT VARIABLES:FILLED	Y=-2.9178	
	ALL VARIABLES:FILLED	Y=-2.9178	

SECOND BASEMAN											
TIME PERIOD 1			**WALD:GLOBAL NULL BETA=0**			**RESIDUAL**			**HOSMER-LEMESHOW**		
		R^2	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**
	SELECT VARIABLES					8	6	0.2381			
	SELECT VARIABLES:FILLED					11	9	0.2757			
	ALL VARIABLES:FILLED					11	9	0.2757			
TIME PERIOD 2	**SELECT VARIABLES**					15	14	0.3782			
	SELECT VARIABLES:FILLED					15	14	0.3782			
	ALL VARIABLES:FILLED					15	14	0.3782			
TIME PERIOD 3	**SELECT VARIABLES**	0.4812	4.169	1	0.0412	9.1871	25	0.9983	3.674	7	0.8165
	SELECT VARIABLES:FILLED	0.4711	6.179	1	0.0129	19.0327	29	0.9204	2.4593	7	0.9301
	ALL VARIABLES:FILLED	0.4711	6.179	1	0.0129	19.0327	29	0.9204	2.4593	7	0.9301
TIME PERIOD 4	**SELECT VARIABLES**					25	24	0.4058			
	SELECT VARIABLES:FILLED					25	24	0.4058			
	ALL VARIABLES:FILLED					25	24	0.4058			

SECOND BASEMAN			
TIME PERIOD 1		**MODEL**	**CLASSIFICATION PERCENT**
	SELECT VARIABLES	Y=-1.0986	
	SELECT VARIABLES:FILLED	Y=-1.5041	
	ALL VARIABLES:FILLED	Y=-1.5041	
TIME PERIOD 2	**SELECT VARIABLES**	Y=-1.0116	
	SELECT VARIABLES:FILLED	Y=-1.0116	
	ALL VARIABLES:FILLED	Y=-1.0116	
TIME PERIOD 3	**SELECT VARIABLES**	Y=0.0113**RBI**-9.2794	87.09677419
	SELECT VARIABLES:FILLED	Y=54.8977**SLUGGINGPERCENT**-22.6661	93.5483871
	ALL VARIABLES:FILLED	Y=54.8977**SLUGGINGPERCENT**-22.6661	93.5483871
TIME PERIOD 4	**SELECT VARIABLES**	Y=-1.9924	
	SELECT VARIABLES:FILLED	Y=-1.9924	
	ALL VARIABLES:FILLED	Y=-1.9924	

THIRD BASEMAN											
TIME PERIOD 1			WALD:GLOBAL NULL BETA=0			RESIDUAL			HOSMER-LEMESHOW		
		R^2	χ^2	DF	P-VALUE	χ^2	DF	P-VALUE	χ^2	DF	P-VALUE
	SELECT VARIABLES					10	8	0.265			
	SELECT VARIABLES:FILLED					13	11	0.2933			
	ALL VARIABLES:FILLED					13	11	0.2933			
TIME PERIOD 2	SELECT VARIABLES					11	10	0.3575			
	SELECT VARIABLES:FILLED					13	12	0.369			
	ALL VARIABLES:FILLED					13	12	0.369			
TIME PERIOD 3	SELECT VARIABLES					19	18	0.3918			
	SELECT VARIABLES:FILLED					26	26	0.4631			
	ALL VARIABLES:FILLED					26	25	0.4076			
TIME PERIOD 4	SELECT VARIABLES					36	35	0.4215			
	SELECT VARIABLES:FILLED					37	36	0.4226			
	ALL VARIABLES:FILLED					37	36	0.4226			

THIRD BASEMAN			
TIME PERIOD 1		**MODEL**	**CLASSIFICATION PERCENT**
	SELECT VARIABLES	Y=-2.1972	
	SELECT VARIABLES:FILLED	Y=-1.7047	
	ALL VARIABLES:FILLED	Y=-1.7047	
TIME PERIOD 2	**SELECT VARIABLES**	Y=-0.9808	
	SELECT VARIABLES:FILLED	Y=-0.8109	
	ALL VARIABLES:FILLED	Y=-0.8109	
TIME PERIOD 3	**SELECT VARIABLES**	Y=-1.0296	
	SELECT VARIABLES:FILLED	Y=-0.9985	
	ALL VARIABLES:FILLED	Y=-0.9985	
TIME PERIOD 4	**SELECT VARIABLES**	Y=-1.8245	
	SELECT VARIABLES:FILLED	Y=-1.8563	
	ALL VARIABLES:FILLED	Y=-1.8563	

SHORT STOP											
TIME PERIOD 1			**WALD:GLOBAL NULL BETA=0**			**RESIDUAL**			**HOSMER-LEMESHOW**		
		R^2	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**
	SELECT VARIABLES					11	8	0.2017			
	SELECT VARIABLES:FILLED					12	9	0.2133			
	ALL VARIABLES:FILLED					12	9	0.2133			
TIME PERIOD 2	**SELECT VARIABLES**					12	11	0.3636			
	SELECT VARIABLES:FILLED					15	14	0.3782			
	ALL VARIABLES:FILLED					15	14	0.3782			
TIME PERIOD 3	**SELECT VARIABLES**					30	29	0.414			
	SELECT VARIABLES:FILLED	0.3568	7.9886	1	0.0047	31.1141	32	0.5112	3.7023	8	0.8829
	ALL VARIABLES:FILLED	0.3568	7.9886	1	0.0047	31.1939	37	0.7374	3.7023	8	0.8829
TIME PERIOD 4	**SELECT VARIABLES**					19	18	0.3918			
	SELECT VARIABLES:FILLED					19	18	0.3918			
	ALL VARIABLES:FILLED					19	18	0.3918			

SHORT STOP			
TIME PERIOD 1		**MODEL**	**CLASSIFICATION PERCENT**
	SELECT VARIABLES	Y=-0.1823	
	SELECT VARIABLES:FILLED	Y=0	
	ALL VARIABLES:FILLED	Y=0	
TIME PERIOD 2	**SELECT VARIABLES**	Y=-0.6931	
	SELECT VARIABLES:FILLED	Y=-0.6931	
	ALL VARIABLES:FILLED	Y=-0.6931	
TIME PERIOD 3	**SELECT VARIABLES**	Y=-1.3863	
	SELECT VARIABLES:FILLED	Y=0.0797**TRIPLES**-6.4497	87.17948718
	ALL VARIABLES:FILLED	Y=0.0797**TRIPLES**-6.4497	87.17948718
TIME PERIOD 4	**SELECT VARIABLES**	Y=-1.3218	
	SELECT VARIABLES:FILLED	Y=-1.3218	
	ALL VARIABLES:FILLED	Y=-1.3218	

LEFT FIELDER											
TIME PERIOD 1			WALD:GLOBAL NULL BETA=0			RESIDUAL			HOSMER-LEMESHOW		
		R^2	χ^2	DF	P-VALUE	χ^2	DF	P-VALUE	χ^2	DF	P-VALUE
	SELECT VARIABLES					6	5	0.3062			
	SELECT VARIABLES:FILLED					6	5	0.3062			
	ALL VARIABLES:FILLED					6	5	0.3062			
TIME PERIOD 2	SELECT VARIABLES					16	15	0.3821			
	SELECT VARIABLES:FILLED					16	15	0.3821			
	ALL VARIABLES:FILLED					16	15	0.3821			
TIME PERIOD 3	SELECT VARIABLES	0.4151	6.4468	1	0.0111	56.6798	28	0.0011	17.4021	8	0.0262
	SELECT VARIABLES:FILLED	0.4254	9.6784	1	0.0019	60.8086	32	0.0016	15.621	7	0.0288
	ALL VARIABLES:FILLED	0.4254	9.6784	1	0.0019	72.35	36	0.0003	15.621	7	0.0288
TIME PERIOD 4	SELECT VARIABLES					33	32	0.418			
	SELECT VARIABLES:FILLED					33	33	0.4673			
	ALL VARIABLES:FILLED					33	33	0.4673			

LEFT FIELDER			
TIME PERIOD 1		**MODEL**	**CLASSIFICATION PERCENT**
	SELECT VARIABLES	Y=1.6094	
	SELECT VARIABLES:FILLED	Y=1.6094	
	ALL VARIABLES:FILLED	Y=1.6094	
TIME PERIOD 2	**SELECT VARIABLES**	Y=-1.4663	
	SELECT VARIABLES:FILLED	Y=-1.4663	
	ALL VARIABLES:FILLED	Y=-1.4663	
TIME PERIOD 3	**SELECT VARIABLES**	Y=0.00493**RUNS**-5.7021	91.89189189
	SELECT VARIABLES:FILLED	Y=0.0138**DOUBLES**-5.3362	84.09090909
	ALL VARIABLES:FILLED	Y=0.0138**DOUBLES**-5.3362	84.09090909
TIME PERIOD 4	**SELECT VARIABLES**	Y=-2.7408	
	SELECT VARIABLES:FILLED	Y=-2.7408	
	ALL VARIABLES:FILLED	Y=-2.7408	

CENTER FIELDER											
TIME PERIOD 1			WALD:GLOBAL NULL BETA=0			RESIDUAL			HOSMER-LEMESHOW		
		R^2	χ^2	DF	P-VALUE	χ^2	DF	P-VALUE	χ^2	DF	P-VALUE
	SELECT VARIABLES					7	6	0.3208			
	SELECT VARIABLES:FILLED					7	6	0.3208			
	ALL VARIABLES:FILLED					7	6	0.3208			
TIME PERIOD 2	SELECT VARIABLES					12	11	0.3636			
	SELECT VARIABLES:FILLED					16	15	0.3821			
	ALL VARIABLES:FILLED					16	15	0.3821			
TIME PERIOD 3	SELECT VARIABLES	0.5321	4.9604	1	0.0259	21.4992	29	0.8403	10.1656	8	0.2536
	SELECT VARIABLES:FILLED					35.0305	32	0.3262			
	ALL VARIABLES:FILLED					37	36	0.4226			
TIME PERIOD 4	SELECT VARIABLES					33	32	0.418			
	SELECT VARIABLES:FILLED					33	33	0.4673			
	ALL VARIABLES:FILLED					33	33	0.4673			

CENTER FIELDER			
TIME PERIOD 1		**MODEL**	**CLASSIFICATION PERCENT**
	SELECT VARIABLES	Y=-0.9163	
	SELECT VARIABLES:FILLED	Y=-0.9163	
	ALL VARIABLES:FILLED	Y=-0.9163	
TIME PERIOD 2	**SELECT VARIABLES**	Y=0	
	SELECT VARIABLES:FILLED	Y=0.5108	
	ALL VARIABLES:FILLED	Y=0.5108	
TIME PERIOD 3	**SELECT VARIABLES**	Y=43.0637**OPS**-36.4753	86.48648649
	SELECT VARIABLES:FILLED	Y=-0.9933	
	ALL VARIABLES:FILLED	Y=-0.9933	
TIME PERIOD 4	**SELECT VARIABLES**	Y=-3.4657	
	SELECT VARIABLES:FILLED	Y=-3.4657	
	ALL VARIABLES:FILLED	Y=-3.4657	

RIGHT FIELDER											
TIME PERIOD 1			**WALD:GLOBAL NULL BETA=0**			**RESIDUAL**			**HOSMER-LEMESHOW**		
		R^2	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**	**χ^2**	**DF**	**P-VALUE**
	SELECT VARIABLES										
	SELECT VARIABLES:FILLED										
	ALL VARIABLES:FILLED										
TIME PERIOD 2	**SELECT VARIABLES**					13	12	0.369			
	SELECT VARIABLES:FILLED					13	12	0.369			
	ALL VARIABLES:FILLED					13	12	0.369			
TIME PERIOD 3	**SELECT VARIABLES**					23	22	0.4017			
	SELECT VARIABLES:FILLED					33	32	0.418			
	ALL VARIABLES:FILLED					33	32	0.418			
TIME PERIOD 4	**SELECT VARIABLES**	0.2844	5.0816	1	0.0242	11.4132	36	1	0.5249	8	0.9998
	SELECT VARIABLES:FILLED					41.4281	37	0.2835			
	ALL VARIABLES:FILLED					41.7464	40	0.3948			

RIGHT FIELD			
TIME PERIOD 1		**MODEL**	**CLASSIFICATION PERCENT**
	SELECT VARIABLES		
	SELECT VARIABLES:FILLED		
	ALL VARIABLES:FILLED		
TIME PERIOD 2	**SELECT VARIABLES**	Y=-0.1542	
	SELECT VARIABLES:FILLED	Y=-0.1542	
	ALL VARIABLES:FILLED	Y=-0.1542	
TIME PERIOD 3	**SELECT VARIABLES**	Y=-1.2809	
	SELECT VARIABLES:FILLED	Y=-1.1394	
	ALL VARIABLES:FILLED	Y=-1.1394	
TIME PERIOD 4	**SELECT VARIABLES**	Y=0.0555**INTENTIONALWALKS**-9.4257	88.0952381
	SELECT VARIABLES:FILLED	Y=-2.2513	
	ALL VARIABLES:FILLED	Y=-2.2513	

APPENDIX B

THE PROGRAM

THE PROGRAM

```
/*************************************************************/
/*                          KELSEY ZEMLER                    */
/*                         THESIS FALL 2013                  */
/*************************************************************/

/*READING IN THE MASTER DATA SET FROM EXCEL*/

PROC IMPORT
DATAFILE='F:\Thesis-Baseball\MASTER2_NEW.xlsx'
OUT=MAINSET
DBMS=XLSX
REPLACE
;

/*CREATING A FORMAT*/

PROC FORMAT;
VALUE $HOF
'Y'='1'
'N'='0'
;
/*SORTING THE DATA SET BY POSITION AND YEAR INDUCTED/FIRST NOMINATED*/
PROC SORT DATA=MAINSET;
      BY POSITION YEAR_1ST_NOMINATED_or_INDUCTED;

RUN;
/*CREATING 9 DIFFERENT DATA SETS*/

DATA FIRST
       SECOND
       THIRD
       CATCHER
```

```
        PITCHER
        LEFT
        RIGHT
        CENTER
        SHORT;
SET MAINSET;
CS1=CS*1;
DROP CS;

FORMAT HOF $HOF.;
IF POSITION='1' THEN OUTPUT PITCHER;
IF POSITION='2' THEN OUTPUT CATCHER;
IF POSITION='3' THEN OUTPUT FIRST;
IF POSITION='4' THEN OUTPUT SECOND;
IF POSITION='5' THEN OUTPUT THIRD;
IF POSITION='6' THEN OUTPUT SHORT;
IF POSITION='7' THEN OUTPUT LEFT;
IF POSITION='8' THEN OUTPUT CENTER;
IF POSITION='9' THEN OUTPUT RIGHT;

RUN;

/*RUNNING LOGISITIC REGRESSION FOR PITCHERS*/

/*TO DETERMINE THE PREDICTING VARIABLES, I LOOKED AT THE PITCHER'S DATA
SET
AND CHOSE THE VARIABLES THAT HAD THE LEAST AMOUNT OF MISSING DATA. THE
FOLLOWING
VARIABLES WERE PICKED BASED ON THIS JUDGEMENT*/

ODS RTF FILE='PITCHER.RTF';/*THIS OUTPUTS THE RESULTS INTO A WORD
DOCUMENT*/

PROC LOGISTIC DATA=PITCHER OUTEST=EST_P COVOUT;
```

```
TITLE'LOGISTIC REGRESSION FOR PITCHERS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES
HOME_RUNS RBIS STOLEN_BASES WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__
SLUGGING__
OPS OPS_ TOTAL_BASES HIT_BY_PITCH G Inn Ch PO A E DP Fld_  RF_9 RF_G
W L W_L_ ERA G_1 GS_1 GF CG_1 SHO SV IP H R ER HR BB SO HBP BK WP_1 BF
ERA_
WHIP H_9 HR_9 BB_9 SO_9 SO_BB

                                          /SELECTION=STEPWISE/*THE
METHOD OF ANALYSIS*/
                                           CTABLE/*CLASSIFICATION
TABLE*/
                                           RSQUARE/*COX AND SNELL
STATISTIC*/
                                           LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                           OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL /*CREATES A DATA SET WITH PROBABILITIES

                                       AND CONFIDENCE INTERVALS*/
                                            ;
RUN;

ODS RTF CLOSE;
/*SINCE THE HOSMER LEMESHOW STATISTIC CAME BACK AS SIGNIFICANT, THE
REGRESSION WAS RE-RUN USING ONLY THE VARIABLES IDENTIFIED
BY THE STEPWISE MODEL. THEN USING THE ENTER METHOD, THE REGRESSION WAS
RE-RUN*/
ODS RTF FILE='PITCHER1.RTF';

PROC LOGISTIC DATA=PITCHER OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR PITCHERS:ENTER';
```

```
MODEL HOF(EVENT='1')= AT_BATS TRIPLES W GS_1 CG_1 SV WP_1

                                          /SELECTION=FORWARD
                                           CTABLE
                                           RSQUARE
                                           LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                           OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                           ;
RUN;

ODS RTF CLOSE;

/*RUNNING LOGISITIC REGRESSION FOR CATCHERS*/

/*TO DETERMINE THE PREDICTING VARIABLES, I LOOKED AT THE CATCHER'S DATA
SET
AND CHOSE THE VARIABLES THAT HAD THE LEAST AMOUNT OF MISSING DATA. THE
FOLLOWING
VARIABLES WERE PICKED BASED ON THIS JUDGEMENT*/

ODS RTF FILE='CATCHER.RTF';

PROC LOGISTIC DATA=CATCHER OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CATCHERS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES CAUGHT_STEALING WALKS STRIKE_OUTS BATTING_AVERAGE
ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G Ch PO A E DP Fld_ RF_G PB WP SB

                                          /SELECTION=STEPWISE
```

```
                                                CTABLE
                                                RSQUARE
                                                LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;

ODS RTF CLOSE;
/*RUNNING LOGISITIC REGRESSION FOR FIRST BASEMEN*/

/*TO DETERMINE THE PREDICTING VARIABLES, I LOOKED AT THE FIRST BASEMAN
DATA SET
AND CHOSE THE VARIABLES THAT HAD THE LEAST AMOUNT OF MISSING DATA. THE
FOLLOWING
VARIABLES WERE PICKED BASED ON THIS JUDGEMENT*/
ODS RTF FILE='FIRST.RTF';

PROC LOGISTIC DATA=FIRST OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR FIRST BASEMEN';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES CAUGHT_STEALING WALKS STRIKE_OUTS BATTING_AVERAGE
ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG  Ch PO A E DP Fld_ RF_G

                                               /SELECTION=STEPWISE
                                                CTABLE
                                                RSQUARE
                                                LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
```

```
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;

ODS RTF CLOSE;
/*RUNNING LOGISITIC REGRESSION FOR SECOND BASEMEN*/

/*TO DETERMINE THE PREDICTING VARIABLES, I LOOKED AT THE SECOND BASEMEN
DATA SET
AND CHOSE THE VARIABLES THAT HAD THE LEAST AMOUNT OF MISSING DATA. THE
FOLLOWING
VARIABLES WERE PICKED BASED ON THIS JUDGEMENT*/
ODS RTF FILE='SECOND.RTF';

PROC LOGISTIC DATA=SECOND OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SECOND BASEMEN';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG  Ch PO A E DP Fld_ RF_G

                                               /SELECTION=STEPWISE
                                                CTABLE
                                                RSQUARE
                                                LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;

ODS RTF CLOSE;
```

```
/*RUNNING LOGISITIC REGRESSION FOR THIRD BASEMEN*/

/*TO DETERMINE THE PREDICTING VARIABLES, I LOOKED AT THE THIRD BASEMEN
DATA SET
AND CHOSE THE VARIABLES THAT HAD THE LEAST AMOUNT OF MISSING DATA. THE
FOLLOWING
VARIABLES WERE PICKED BASED ON THIS JUDGEMENT*/
ODS RTF FILE='THIRD.RTF';

PROC LOGISTIC DATA=THIRD OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR THIRD BASEMEN';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G Ch PO A E DP Fld_  RF_G

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;
/*RUNNING LOGISITIC REGRESSION FOR SHORT STOPS*/

/*TO DETERMINE THE PREDICTING VARIABLES, I LOOKED AT THE SHORT STOP'S
DATA SET
AND CHOSE THE VARIABLES THAT HAD THE LEAST AMOUNT OF MISSING DATA. THE
FOLLOWING
```

```
VARIABLES WERE PICKED BASED ON THIS JUDGEMENT*/
ODS RTF FILE='SHORT.RTF';

PROC LOGISTIC DATA=SHORT OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SHORT STOPS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G Ch PO A E DP Fld_  RF_G

                                          /SELECTION=STEPWISE
                                           CTABLE
                                           RSQUARE
                                           LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                           OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                           ;
RUN;

ODS RTF CLOSE;
/*RUNNING LOGISITIC REGRESSION FOR LEFT FIELDERS*/

/*TO DETERMINE THE PREDICTING VARIABLES, I LOOKED AT THE LEFT FIELDER
DATA SET
AND CHOSE THE VARIABLES THAT HAD THE LEAST AMOUNT OF MISSING DATA. THE
FOLLOWING
VARIABLES WERE PICKED BASED ON THIS JUDGEMENT*/
ODS RTF FILE='LEFT.RTF';

PROC LOGISTIC DATA=LEFT OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR LEFT FIELDERS';
```

```
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
CAUGHT_STEALING HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG Ch PO A E DP Fld_  RF_G

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;
/*SINCE THE HOSMER LEMESHOW STATISTIC CAME BACK AS SIGNIFICANT, THE
REGRESSION WAS RE-RUN USING ONLY THE VARIABLES IDENTIFIED
BY THE STEPWISE MODEL. THEN USING THE ENTER METHOD, THE REGRESSION WAS
RE-RUN*/

ODS RTF FILE='LEFT1.RTF';

PROC LOGISTIC DATA=LEFT OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR LEFT FIELDERS:ENTER';
MODEL HOF(EVENT='1')= RUNS

                                        /SELECTION=FORWARD
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
```

```
                                                    OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                    ;
RUN;

ODS RTF CLOSE;
/*RUNNING LOGISITIC REGRESSION FOR CENTER FIELDERS*/

/*TO DETERMINE THE PREDICTING VARIABLES, I LOOKED AT THE CENTER FIELDER
DATA SET
AND CHOSE THE VARIABLES THAT HAD THE LEAST AMOUNT OF MISSING DATA. THE
FOLLOWING
VARIABLES WERE PICKED BASED ON THIS JUDGEMENT*/
ODS RTF FILE='CENTER.RTF';

PROC LOGISTIC DATA=CENTER OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CENTER FIELDERS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
CAUGHT_STEALING HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G Ch PO A E DP Fld_  RF_G

                                                   /SELECTION=STEPWISE
                                                    CTABLE
                                                    RSQUARE
                                                    LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                    OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                    ;
RUN;

ODS RTF CLOSE;
```

```
/*RUNNING LOGISITIC REGRESSION FOR RIGHT FIELDERS*/

/*TO DETERMINE THE PREDICTING VARIABLES, I LOOKED AT THE RIGHT FIELDER
DATA SET
AND CHOSE THE VARIABLES THAT HAD THE LEAST AMOUNT OF MISSING DATA. THE
FOLLOWING
VARIABLES WERE PICKED BASED ON THIS JUDGEMENT*/
ODS RTF FILE='RIGHT.RTF';

PROC LOGISTIC DATA=RIGHT OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR RIGHT FIELDERS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
CAUGHT_STEALING HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG Ch PO A E DP Fld_  RF_G

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;

/*****************************************************************
******/
/*ADDING VARIABLES INTO EACH MODEL THAT HAD A DESCENT AMOUNT OF MISSING
DATA*/
```

```
/*************************************************************************
******/

/*PITCHERS*/
ODS RTF FILE='PITCHER_ADD.RTF';

PROC LOGISTIC DATA=PITCHER OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR PITCHERS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS
CAUGHT_STEALING GDP GS CG RUNS HITS DOUBLES TRIPLES
HOME_RUNS RBIS STOLEN_BASES WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__
SLUGGING__
OPS OPS_ TOTAL_BASES HIT_BY_PITCH G Inn Ch PO A E DP Fld_  RF_9 RF_G
W L W_L_ ERA G_1 GS_1 GF CG_1 SHO SV IP H R ER HR BB SO HBP BK WP_1 BF
ERA_
WHIP H_9 HR_9 BB_9 SO_9 SO_BB

                                              /SELECTION=STEPWISE
                                               CTABLE
                                               RSQUARE
                                               LACKFIT;
                                               OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                               ;
RUN;

ODS RTF CLOSE;

/*CATCHERS*/

ODS RTF FILE='CATCHER_ADD.RTF';

PROC LOGISTIC DATA=CATCHER OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CATCHERS';
```

```
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
GS CG STOLEN_BASES CAUGHT_STEALING WALKS STRIKE_OUTS BATTING_AVERAGE
ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G Ch PO A E DP Fld_ RF_G PB WP SB

                                          /SELECTION=STEPWISE
                                           CTABLE
                                           RSQUARE
                                           LACKFIT;
                                           OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                           ;
RUN;

ODS RTF CLOSE;

/*FIRST BASEMEN*/

ODS RTF FILE='FIRST_ADD.RTF';

PROC LOGISTIC DATA=FIRST OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR FIRST BASEMEN';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES CAUGHT_STEALING WALKS GDP STRIKE_OUTS BATTING_AVERAGE
ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG  Ch PO A E DP Fld_ RF_G

                                          /SELECTION=STEPWISE
                                           CTABLE
                                           RSQUARE
```

```
                                                    LACKFIT;
                                                    OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                    ;
RUN;

ODS RTF CLOSE;

/*SECOND BASEMEN*/

ODS RTF FILE='SECOND_ADD.RTF';

PROC LOGISTIC DATA=SECOND OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SECOND BASEMEN';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES CAUGHT_STEALING GDP STRIKE_OUTS BATTING_AVERAGE ON_BASE__
SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG  Ch PO A E DP Fld_ RF_G

                                                   /SELECTION=STEPWISE
                                                    CTABLE
                                                    RSQUARE
                                                    LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                    OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                    ;
RUN;

ODS RTF CLOSE;

/*THIRD BASEMEN*/
```

```
ODS RTF FILE='THIRD_ADD.RTF';

PROC LOGISTIC DATA=THIRD OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR THIRD BASEMEN';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES CAUGHT_STEALING GDP STRIKE_OUTS BATTING_AVERAGE ON_BASE__
SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G Ch PO A E DP Fld_  RF_G

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;

/*SHORT STOP*/

ODS RTF FILE='SHORT_ADD.RTF';

PROC LOGISTIC DATA=SHORT OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SHORT STOPS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES CAUGHT_STEALING STRIKE_OUTS BATTING_AVERAGE ON_BASE__
SLUGGING__
```

```
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG Ch PO A E DP Fld_  RF_G

                                          /SELECTION=STEPWISE
                                           CTABLE
                                           RSQUARE
                                           LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                           OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                           ;
RUN;

ODS RTF CLOSE;


/*LEFT FIELDER*/

ODS RTF FILE='LEFT_ADD.RTF';

PROC LOGISTIC DATA=LEFT OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR LEFT FIELDERS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
CAUGHT_STEALING HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES CAUGHT_STEALING STRIKE_OUTS BATTING_AVERAGE ON_BASE__
SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG Ch PO A E DP Fld_  RF_G

                                          /SELECTION=STEPWISE
                                           CTABLE
                                           RSQUARE
                                           LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
```

```
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;

ODS RTF CLOSE;
/*CENTER FIELDER*/
ODS RTF FILE='CENTER_ADD.RTF';

PROC LOGISTIC DATA=CENTER OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CENTER FIELDERS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
CAUGHT_STEALING HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES CAUGHT_STEALING STRIKE_OUTS BATTING_AVERAGE ON_BASE__
SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG Ch PO A E DP Fld_  RF_G

                                               /SELECTION=STEPWISE
                                                CTABLE
                                                RSQUARE
                                                LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;

ODS RTF CLOSE;

/*RIGHT FIELDER*/

ODS RTF FILE='RIGHT_ADD.RTF';
```

```
PROC LOGISTIC DATA=RIGHT OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR RIGHT FIELDERS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
CAUGHT_STEALING HITS DOUBLES TRIPLES HOME_RUNS RBIS
GDP STOLEN_BASES STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG Ch PO A E DP Fld_  RF_G

                                          /SELECTION=STEPWISE
                                           CTABLE
                                           RSQUARE
                                           LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                           OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                           ;
RUN;

ODS RTF CLOSE;

/*************************************************/
/*CREATING A NEW DATA SET USING THE VARAIBLE MEAN AS*/
/*A SUBSTITUTION FOR MISSING DATA                  */
/*************************************************/
/*MEAN SUBSTITUTIONS WERE DONE IN SPSS 19. THEN I MADE THEM INTO AN
EXCEL SPREADSHEET.
NEXT, I IMPORTED THE DATA SETS, INDIVIDUALLY, INTO SAS*/
PROC IMPORT
DATAFILE='F:\Thesis-Baseball\PITCHER_FILLED.xlsx'
OUT=PITCHER_FILL
DBMS=XLSX
REPLACE
;
RUN;
```

```
/*RE-RUNNING THE LOGISTIC REGRESSION PRIOR TO THIS*/
ODS RTF FILE='PITCHER_ADD.RTF';

PROC LOGISTIC DATA=PITCHER_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR PITCHERS:ADD AND FILLED';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS
CAUGHT_STEALING GDP GS CG RUNS HITS DOUBLES TRIPLES
HOME_RUNS RBIS STOLEN_BASES WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__
SLUGGING__
OPS OPS_ TOTAL_BASES HIT_BY_PITCH G Inn Ch PO A E DP Fld_  RF_9 RF_G
W L W_L_ ERA G_1 GS_1 GF CG_1 SHO SV IP H R ER HR BB SO HBP BK WP_1 BF
ERA_
WHIP H_9 HR_9 BB_9 SO_9 SO_BB

                                              /SELECTION=STEPWISE
                                               CTABLE
                                               RSQUARE
                                               LACKFIT;
                                               OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                               ;
RUN;

ODS RTF CLOSE;

/*RUNNING THE REGRESSION USING ALL VARIABLES*/
ODS RTF FILE='PITCHER_ALL_VARS.RTF';

PROC LOGISTIC DATA=PITCHER_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR PITCHERS: ADDED AND FILLED';
```

```
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG PB WP SB CS CS_ lgCS_ PO_1
W L W_L_ ERA G_1 GS_1 GF CG_1 SHO SV IP H R ER HR BB IBB SO HBP BK WP_1
BF ERA_ WHIP H_9 HR_9 BB_9 SO_9 SO_BB

                                              /SELECTION=STEPWISE
                                               CTABLE
                                               RSQUARE
                                               LACKFIT;
                                               OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                               ;
RUN;

ODS RTF CLOSE;

/*IMPORTING COMPLETE CATCHER DATA*/
PROC IMPORT
DATAFILE='F:\Thesis-Baseball\catcher_FILLED.xlsx'
OUT=CATCHER_FILL
DBMS=XLSX
REPLACE
;
RUN;
/*RE-RUNNING USING SAME VARIABLES AS BEFORE*/
ODS RTF FILE='CATCHER_ADD_F.RTF';

PROC LOGISTIC DATA=CATCHER_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CATCHERS:ADDED AND FILLED';
```

```
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
GS CG STOLEN_BASES CAUGHT_STEALING WALKS STRIKE_OUTS BATTING_AVERAGE
ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G Ch PO A E DP Fld_ RF_G PB WP SB

                                          /SELECTION=STEPWISE
                                           CTABLE
                                           RSQUARE
                                           LACKFIT;
                                           OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                           ;
RUN;

ODS RTF CLOSE;

/*RUNNING USING ALL VARIABLES*/
ODS RTF FILE='CATCHER_FILLED.RTF';

PROC LOGISTIC DATA=CATCHER_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CATCHERS:FILLED. ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG PB WP SB CS CS_ lgCS_ PO_1

                                          /SELECTION=STEPWISE
                                           CTABLE
                                           RSQUARE
```

```
                                                    LACKFIT;
                                                    OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                    ;
RUN;

ODS RTF CLOSE;

/*IMPORTING FIRST BASE FILLED DATA*/
PROC IMPORT
DATAFILE='F:\Thesis-Baseball\FIRST_FILLED.xlsx'
OUT=FIRST_FILL
DBMS=XLSX
REPLACE
;
RUN;

/*RE-RUNNING USING SAME VARIABLES USING */

ODS RTF FILE='FIRST_ADD_F.RTF';

PROC LOGISTIC DATA=FIRST_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR FIRST BASEMEN:ADDED AND FILLED';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES CAUGHT_STEALING WALKS GDP STRIKE_OUTS BATTING_AVERAGE
ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG  Ch PO A E DP Fld_ RF_G

                                                   /SELECTION=STEPWISE
                                                    CTABLE
                                                    RSQUARE
                                                    LACKFIT;
```

```
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;

ODS RTF CLOSE;

/*RUNNING USING ALL VARIABLES*/
ODS RTF FILE='FIRST_FILLED.RTF';

PROC LOGISTIC DATA=FIRST_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR FIRST:FILLED. ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                               /SELECTION=STEPWISE
                                                CTABLE
                                                RSQUARE
                                                LACKFIT;
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;

ODS RTF CLOSE;

/*SECOND BASEMEN*/

/*IMPORTING DATA*/
```

```
PROC IMPORT
DATAFILE='F:\Thesis-Baseball\SECOND_FILLED.xlsx'
OUT=SECOND_FILL
DBMS=XLSX
REPLACE
;
RUN;

/*RERUNNING USING SAME VARIABLES*/
ODS RTF FILE='SECOND_ADD_F.RTF';

PROC LOGISTIC DATA=SECOND_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SECOND BASEMEN';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES CAUGHT_STEALING GDP STRIKE_OUTS BATTING_AVERAGE ON_BASE__
SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG  Ch PO A E DP Fld_ RF_G

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;

/*RERUNNING USING ALL VARIABLES*/
```

```
ODS RTF FILE='SECOND_FILLED.RTF';

PROC LOGISTIC DATA=SECOND_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SECOND:FILLED. ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                              /SELECTION=STEPWISE
                                               CTABLE
                                               RSQUARE
                                               LACKFIT;
                                               OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                               ;
RUN;

ODS RTF CLOSE;
/*THIRD*/

/*IMPORT DATA*/
PROC IMPORT
DATAFILE='F:\Thesis-Baseball\THIRD_FILLED.xlsx'
OUT=THIRD_FILL
DBMS=XLSX
REPLACE
;
RUN;
/*RERUNNING USING SAME VARIABLES*/
ODS RTF FILE='THIRD_ADD_F.RTF';
```

```
PROC LOGISTIC DATA=THIRD_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR THIRD BASEMEN:ADDED AND FILLED';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES CAUGHT_STEALING GDP STRIKE_OUTS BATTING_AVERAGE ON_BASE__
SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG  Ch PO A E DP Fld_ RF_G

                                          /SELECTION=STEPWISE
                                           CTABLE
                                           RSQUARE
                                           LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                           OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                           ;
RUN;

ODS RTF CLOSE;
/*RERUNNING USING ALL VARIABLES*/

ODS RTF FILE='THIRD_FILLED.RTF';

PROC LOGISTIC DATA=THIRD_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR THIRD:FILLED. ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG
```

```
                                                    /SELECTION=STEPWISE
                                                     CTABLE
                                                     RSQUARE
                                                     LACKFIT;
                                                     OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                     ;
RUN;

ODS RTF CLOSE;

/*SHORT STOP*/

/*IMPORT DATA*/
PROC IMPORT
DATAFILE='F:\Thesis-Baseball\SHORT_FILLED.xlsx'
OUT=SHORT_FILL
DBMS=XLSX
REPLACE
;
RUN;
/*RERUNNING USING SAME VARIABLES BUT FILLED*/
ODS RTF FILE='SHORT_ADD_F.RTF';

PROC LOGISTIC DATA=SHORT_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SHORT STOPS:ADDED AND FILLED';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES CAUGHT_STEALING STRIKE_OUTS BATTING_AVERAGE ON_BASE__
SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG Ch PO A E DP Fld_  RF_G
```

```
                                                /SELECTION=STEPWISE
                                                 CTABLE
                                                 RSQUARE
                                                 LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                 OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                 ;
RUN;

ODS RTF CLOSE;

/*RERUNNING USING ALL VARIABLES*/

ODS RTF FILE='SHORT_FILLED.RTF';

PROC LOGISTIC DATA=SHORT_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SHORT STOP:FILLED. ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                                /SELECTION=STEPWISE
                                                 CTABLE
                                                 RSQUARE
                                                 LACKFIT;
                                                 OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                 ;
RUN;
```

```
ODS RTF CLOSE;

/*LEFT FIELD*/

/*IMPORT DATA*/
PROC IMPORT
DATAFILE='F:\Thesis-Baseball\LEFT_FILLED.xlsx'
OUT=LEFT_FILL
DBMS=XLSX
REPLACE
;
RUN;

/*RE-RUNNING USING FILLED DATA*/

ODS RTF FILE='LEFT_ADD_F.RTF';

PROC LOGISTIC DATA=LEFT_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR LEFT FIELDERS:ADDED AND FILLED';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
CAUGHT_STEALING HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES CAUGHT_STEALING STRIKE_OUTS BATTING_AVERAGE ON_BASE__
SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG Ch PO A E DP Fld_  RF_G

                                          /SELECTION=STEPWISE
                                           CTABLE
                                           RSQUARE
                                           LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                           OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
```

```
                                                    ;
RUN;

ODS RTF CLOSE;

/*RERUNNING USING ALL VARIABLES*/

ODS RTF FILE='LEFT_FILLED.RTF';

PROC LOGISTIC DATA=LEFT_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR LEFT FIELDERS:FILLED. ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                                   /SELECTION=STEPWISE
                                                    CTABLE
                                                    RSQUARE
                                                    LACKFIT;
                                                    OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                    ;
RUN;

ODS RTF CLOSE;

/*CENTER FIELD*/

/*IMPORT DATA*/
PROC IMPORT
```

```
DATAFILE='F:\Thesis-Baseball\CENTER_FILLED.xlsx'
OUT=CENTER_FILL
DBMS=XLSX
REPLACE
;
RUN;

/*RE-RUNNING USING FILLED DATA*/

ODS RTF FILE='CENTER_ADD_F.RTF';

PROC LOGISTIC DATA=CENTER_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CENTER FIELDERS:ADDED AND FILLED';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
CAUGHT_STEALING HITS DOUBLES TRIPLES HOME_RUNS RBIS
STOLEN_BASES CAUGHT_STEALING STRIKE_OUTS BATTING_AVERAGE ON_BASE__
SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG Ch PO A E DP Fld_  RF_G

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;

/*RERUNNING USING ALL VARIABLES*/
```

```
ODS RTF FILE='CENTER_FILLED.RTF';

PROC LOGISTIC DATA=CENTER_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CENTER FIELDERS:FILLED. ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;

/*RIGHT FIELD*/

/*IMPORT DATA*/
PROC IMPORT
DATAFILE='F:\Thesis-Baseball\RIGHT_FILLED.xlsx'
OUT=RIGHT_FILL
DBMS=XLSX
REPLACE
;
RUN;
```

```
/*RE-RUNNING USING FILLED DATA*/

ODS RTF FILE='RIGHT_ADD_F.RTF';

PROC LOGISTIC DATA=RIGHT_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR RIGHT FIELDERS:ADDED AND FILLED';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
CAUGHT_STEALING HITS DOUBLES TRIPLES HOME_RUNS RBIS
GDP STOLEN_BASES STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES  HIT_BY_PITCH SACRIFICE_HITS
G GS CG Ch PO A E DP Fld_  RF_G

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;

/*RERUNNING USING ALL VARIABLES*/

ODS RTF FILE='RIGHT_FILLED.RTF';

PROC LOGISTIC DATA=RIGHT_FILL OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR RIGHT FIELDERS:FILLED. ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
```

```
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                                /SELECTION=STEPWISE
                                                 CTABLE
                                                 RSQUARE
                                                 LACKFIT;
                                                 OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                 ;
RUN;


ODS RTF CLOSE;

/*****************************************************/
/*                TIME FRAME ANALYSIS                */
/*****************************************************/

/*SPLITTING EACH OF THE 9 POSITION DATA SETS INTO
TIME FRAMES*/

/*PITCHER*/
DATA PITCHER_T1
       PITCHER_T2
       PITCHER_T3
       PITCHER_T4;
  SET PITCHER_FILL;

IF TIME_PERIOD_NUMBER ='1' THEN OUTPUT PITCHER_T1;
IF TIME_PERIOD_NUMBER ='2' THEN OUTPUT PITCHER_T2;
IF TIME_PERIOD_NUMBER ='3' THEN OUTPUT PITCHER_T3;
```

```
IF TIME_PERIOD_NUMBER ='4' THEN OUTPUT PITCHER_T4;

RUN;
/*CATCHER*/
DATA CATCHER_T1
       CATCHER_T2
       CATCHER_T3
       CATCHER_T4;
  SET CATCHER_FILL;

IF TIME_PERIOD_NUMBER ='1' THEN OUTPUT CATCHER_T1;
IF TIME_PERIOD_NUMBER ='2' THEN OUTPUT CATCHER_T2;
IF TIME_PERIOD_NUMBER ='3' THEN OUTPUT CATCHER_T3;
IF TIME_PERIOD_NUMBER ='4' THEN OUTPUT CATCHER_T4;

RUN;
/*FIRST BASE*/
DATA FIRST_T1
       FIRST_T2
       FIRST_T3
       FIRST_T4;
  SET FIRST_FILL;

IF TIME_PERIOD_NUMBER ='1' THEN OUTPUT FIRST_T1;
IF TIME_PERIOD_NUMBER ='2' THEN OUTPUT FIRST_T2;
IF TIME_PERIOD_NUMBER ='3' THEN OUTPUT FIRST_T3;
IF TIME_PERIOD_NUMBER ='4' THEN OUTPUT FIRST_T4;

RUN;

/*SECOND BASE*/
DATA SECOND_T1
       SECOND_T2
       SECOND_T3
```

```
       SECOND_T4;
  SET SECOND_FILL;

IF TIME_PERIOD_NUMBER ='1' THEN OUTPUT SECOND_T1;
IF TIME_PERIOD_NUMBER ='2' THEN OUTPUT SECOND_T2;
IF TIME_PERIOD_NUMBER ='3' THEN OUTPUT SECOND_T3;
IF TIME_PERIOD_NUMBER ='4' THEN OUTPUT SECOND_T4;

RUN;
/*THIRD BASE*/

DATA THIRD_T1
       THIRD_T2
       THIRD_T3
       THIRD_T4;
  SET THIRD_FILL;

IF TIME_PERIOD_NUMBER ='1' THEN OUTPUT THIRD_T1;
IF TIME_PERIOD_NUMBER ='2' THEN OUTPUT THIRD_T2;
IF TIME_PERIOD_NUMBER ='3' THEN OUTPUT THIRD_T3;
IF TIME_PERIOD_NUMBER ='4' THEN OUTPUT THIRD_T4;

RUN;

/*SHORT STOP*/
DATA SHORT_T1
       SHORT_T2
       SHORT_T3
       SHORT_T4;
  SET SHORT_FILL;

IF TIME_PERIOD_NUMBER ='1' THEN OUTPUT SHORT_T1;
IF TIME_PERIOD_NUMBER ='2' THEN OUTPUT SHORT_T2;
IF TIME_PERIOD_NUMBER ='3' THEN OUTPUT SHORT_T3;
```

```
IF TIME_PERIOD_NUMBER ='4' THEN OUTPUT SHORT_T4;

RUN;
/*LEFT FIELD*/
DATA LEFT_T1
       LEFT_T2
       LEFT_T3
       LEFT_T4;
  SET LEFT_FILL;

IF TIME_PERIOD_NUMBER ='1' THEN OUTPUT LEFT_T1;
IF TIME_PERIOD_NUMBER ='2' THEN OUTPUT LEFT_T2;
IF TIME_PERIOD_NUMBER ='3' THEN OUTPUT LEFT_T3;
IF TIME_PERIOD_NUMBER ='4' THEN OUTPUT LEFT_T4;

RUN;
/*CENTER FIELD*/
DATA CENTER_T1
       CENTER_T2
       CENTER_T3
       CENTER_T4;
  SET CENTER_FILL;

IF TIME_PERIOD_NUMBER ='1' THEN OUTPUT CENTER_T1;
IF TIME_PERIOD_NUMBER ='2' THEN OUTPUT CENTER_T2;
IF TIME_PERIOD_NUMBER ='3' THEN OUTPUT CENTER_T3;
IF TIME_PERIOD_NUMBER ='4' THEN OUTPUT CENTER_T4;

RUN;
/*RIGHT FIELD*/
DATA RIGHT_T1
       RIGHT_T2
       RIGHT_T3
       RIGHT_T4;
```

```
  SET RIGHT_FILL;

IF TIME_PERIOD_NUMBER ='1' THEN OUTPUT RIGHT_T1;
IF TIME_PERIOD_NUMBER ='2' THEN OUTPUT RIGHT_T2;
IF TIME_PERIOD_NUMBER ='3' THEN OUTPUT RIGHT_T3;
IF TIME_PERIOD_NUMBER ='4' THEN OUTPUT RIGHT_T4;

RUN;

/*RUNNING ANALYSIS ON PITCHERS FOR EACH TIME FRAME*/
/*TIME FRAME NUMBER 1*/

ODS RTF FILE='PITCHER1.RTF';

PROC LOGISTIC DATA=PITCHER_T1 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR PITCHERS:FILLED ALL VARS:TIME PERIOD 1';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG PB WP SB CS CS_ lgCS_ PO_1
W L W_L_ ERA G_1 GS_1 GF CG_1 SHO SV IP H R ER HR BB IBB SO HBP BK WP_1
BF ERA_ WHIP H_9 HR_9 BB_9 SO_9 SO_BB
                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;
```

```
ODS RTF CLOSE;
/*TIME FRAME NUMBER 2*/
ODS RTF FILE='PITCHER2.RTF';

PROC LOGISTIC DATA=PITCHER_T2 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR PITCHERS:FILLED ALL VARS:TIME PERIOD 2';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG PB WP SB CS CS_ lgCS_ PO_1
W L W_L_ ERA G_1 GS_1 GF CG_1 SHO SV IP H R ER HR BB IBB SO HBP BK WP_1
BF ERA_ WHIP H_9 HR_9 BB_9 SO_9 SO_BB
                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;


ODS RTF CLOSE;
/*TIME FRAME NUMBER 3*/
ODS RTF FILE='PITCHER3.RTF';

PROC LOGISTIC DATA=PITCHER_T3 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR PITCHERS:FILLED ALL VARS:TIME PERIOD 3';
```

```
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG PB WP SB CS CS_ lgCS_ PO_1
W L W_L_ ERA G_1 GS_1 GF CG_1 SHO SV IP H R ER HR BB IBB SO HBP BK WP_1
BF ERA_ WHIP H_9 HR_9 BB_9 SO_9 SO_BB
                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;


ODS RTF CLOSE;
/*TIME FRAME NUMBER 4*/
ODS RTF FILE='PITCHER4.RTF';

PROC LOGISTIC DATA=PITCHER_T4 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR PITCHERS:FILLED ALL VARS:TIME PERIOD 4';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG PB WP SB CS CS_ lgCS_ PO_1
W L W_L_ ERA G_1 GS_1 GF CG_1 SHO SV IP H R ER HR BB IBB SO HBP BK WP_1
BF ERA_ WHIP H_9 HR_9 BB_9 SO_9 SO_BB
```

```
                                                    /SELECTION=STEPWISE
                                                     CTABLE
                                                     RSQUARE
                                                     LACKFIT;
                                                     OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                     ;
RUN;


ODS RTF CLOSE;

/*RUNNING ANALYSIS ON CATCHERS FOR EACH TIME FRAME*/
/*TIME FRAME NUMBER 1*/

ODS RTF FILE='CATCHER1.RTF';

PROC LOGISTIC DATA=CATCHER_T1 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CATCHERS:FILLED ALL VARS:TIME PERIOD 1';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG PB WP SB CS CS_ lgCS_ PO_1

                                                    /SELECTION=STEPWISE
                                                     CTABLE
                                                     RSQUARE
                                                     LACKFIT;
                                                     OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                     ;
```

```
RUN;


ODS RTF CLOSE;
/*TIME FRAME NUMBER 2*/

ODS RTF FILE='CATCHER2.RTF';

PROC LOGISTIC DATA=CATCHER_T2 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CATCHERS:FILLED ALL VARS:TIME PERIOD 2';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG PB WP SB CS CS_ lgCS_ PO_1

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 3*/
ODS RTF FILE='CATCHER3.RTF';

PROC LOGISTIC DATA=CATCHER_T3 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CATCHERS:FILLED ALL VARS:TIME PERIOD 3';
```

```
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG PB WP SB CS CS_ lgCS_ PO_1

                                          /SELECTION=STEPWISE
                                           CTABLE
                                           RSQUARE
                                           LACKFIT;
                                           OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                           ;
RUN;


ODS RTF CLOSE;
/*TIME FRAME NUMBER 4*/
ODS RTF FILE='CATCHER4.RTF';

PROC LOGISTIC DATA=CATCHER_T4 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CATCHERSS:FILLED ALL VARS:TIME PERIOD 4';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING WALKS
STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG PB WP SB CS CS_ lgCS_ PO_1

                                          /SELECTION=STEPWISE
                                           CTABLE
```

```
                                                RSQUARE
                                                LACKFIT;
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;


ODS RTF CLOSE;
/*RUNNING ANALYSIS ON FIRST BASE FOR EACH TIME FRAME*/

/*TIME FRAME NUMBER 1*/
ODS RTF FILE='FIRST1.RTF';

PROC LOGISTIC DATA=FIRST_T1 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR FIRST:FILLED:ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                               /SELECTION=STEPWISE
                                                CTABLE
                                                RSQUARE
                                                LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;
```

```
ODS RTF CLOSE;
/*TIME FRAME NUMBER 2*/
ODS RTF FILE='FIRST2.RTF';

PROC LOGISTIC DATA=FIRST_T2 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR FIRST:FILLED ALL VARS: TIME 2';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                              /SELECTION=STEPWISE
                                               CTABLE
                                               RSQUARE
                                               LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                               OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                               ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 3*/
ODS RTF FILE='FIRST3.RTF';

PROC LOGISTIC DATA=FIRST_T3 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR FIRST:FILLED ALL VARS:TIME 3';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
```

```
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 4*/
ODS RTF FILE='FIRST4.RTF';

PROC LOGISTIC DATA=FIRST_T4 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR FIRST:FILLED ALL VARS:TIME 4';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
```

```
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;

ODS RTF CLOSE;

/*RUNNING ANALYSIS ON SECOND BASE FOR EACH TIME FRAME*/

/*TIME FRAME NUMBER 1*/
ODS RTF FILE='SECOND1.RTF';

PROC LOGISTIC DATA=SECOND_T1 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SECOND:FILLED:ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                               /SELECTION=STEPWISE
                                                CTABLE
                                                RSQUARE
                                                LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 2*/
```

```
ODS RTF FILE='SECOND2.RTF';

PROC LOGISTIC DATA=SECOND_T2 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SECOND:FILLED ALL VARS: TIME 2';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 3*/
ODS RTF FILE='SECOND3.RTF';

PROC LOGISTIC DATA=SECOND_T3 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SECOND:FILLED ALL VARS:TIME 3';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
```

```
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                                /SELECTION=STEPWISE
                                                 CTABLE
                                                 RSQUARE
                                                 LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                 OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                 ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 4*/
ODS RTF FILE='SECOND4.RTF';

PROC LOGISTIC DATA=SECOND_T4 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SECOND:FILLED ALL VARS:TIME 4';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                                /SELECTION=STEPWISE
                                                 CTABLE
                                                 RSQUARE
                                                 LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                 OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
```

```
                                                    ;
RUN;

ODS RTF CLOSE;

/*RUNNING ANALYSIS ON THIRD BASE FOR EACH TIME FRAME*/

/*TIME FRAME NUMBER 1*/
ODS RTF FILE='THIRD1.RTF';

PROC LOGISTIC DATA=THIRD_T1 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR THIRD:FILLED:ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                                   /SELECTION=STEPWISE
                                                    CTABLE
                                                    RSQUARE
                                                    LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                    OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                    ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 2*/
ODS RTF FILE='THIRD2.RTF';
```

```
PROC LOGISTIC DATA=THIRD_T2 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR THIRD:FILLED ALL VARS: TIME 2';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                          /SELECTION=STEPWISE
                                           CTABLE
                                           RSQUARE
                                           LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                           OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                           ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 3*/
ODS RTF FILE='THIRD3.RTF';

PROC LOGISTIC DATA=THIRD_T3 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR THIRD:FILLED ALL VARS:TIME 3';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG
```

```
                                                /SELECTION=STEPWISE
                                                 CTABLE
                                                 RSQUARE
                                                 LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                 OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                 ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 4*/
ODS RTF FILE='THIRD4.RTF';

PROC LOGISTIC DATA=THIRD_T4 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR THIRD:FILLED ALL VARS:TIME 4';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                                /SELECTION=STEPWISE
                                                 CTABLE
                                                 RSQUARE
                                                 LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                 OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                 ;
RUN;
```

```
ODS RTF CLOSE;
/*RUNNING ANALYSIS ON SHORT STOP FOR EACH TIME FRAME*/

/*TIME FRAME NUMBER 1*/
ODS RTF FILE='SHORT1.RTF';

PROC LOGISTIC DATA=SHORT_T1 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SHORT STOP:FILLED:ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 2*/
ODS RTF FILE='SHORT2.RTF';

PROC LOGISTIC DATA=SHORT_T2 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SHORT STOP:FILLED ALL VARS: TIME 2';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
```

```
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                                /SELECTION=STEPWISE
                                                 CTABLE
                                                 RSQUARE
                                                 LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                 OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                 ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 3*/
ODS RTF FILE='SHORT3.RTF';

PROC LOGISTIC DATA=SHORT_T3 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SHORT STOP:FILLED ALL VARS:TIME 3';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                                /SELECTION=STEPWISE
                                                 CTABLE
                                                 RSQUARE
```

```
                                                LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 4*/
ODS RTF FILE='SHORT4.RTF';

PROC LOGISTIC DATA=SHORT_T4 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR SHORT STOP:FILLED ALL VARS:TIME 4';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                               /SELECTION=STEPWISE
                                                CTABLE
                                                RSQUARE
                                                LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;

ODS RTF CLOSE;

/*RUNNING ANALYSIS ON LEFT FIELD FOR EACH TIME FRAME*/
```

```
/*TIME FRAME NUMBER 1*/
ODS RTF FILE='LEFT1.RTF';

PROC LOGISTIC DATA=LEFT_T1 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR LEFT:FILLED:ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                                /SELECTION=STEPWISE
                                                 CTABLE
                                                 RSQUARE
                                                 LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                 OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                 ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 2*/
ODS RTF FILE='LEFT2.RTF';

PROC LOGISTIC DATA=LEFT_T2 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR LEFT:FILLED ALL VARS: TIME 2';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
```

```
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 3*/
ODS RTF FILE='LEFT3.RTF';

PROC LOGISTIC DATA=LEFT_T3 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR LEFT:FILLED ALL VARS:TIME 3';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
```

```
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 4*/
ODS RTF FILE='LEFT4.RTF';

PROC LOGISTIC DATA=LEFT_T4 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR LEFT:FILLED ALL VARS:TIME 4';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                               /SELECTION=STEPWISE
                                                CTABLE
                                                RSQUARE
                                                LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                ;
RUN;

ODS RTF CLOSE;

/*RUNNING ANALYSIS ON CENTER FIELD FOR EACH TIME FRAME*/

/*TIME FRAME NUMBER 1*/
```

```
ODS RTF FILE='CENTER1.RTF';

PROC LOGISTIC DATA=CENTER_T1 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CENTER:FILLED:ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 2*/
ODS RTF FILE='CENTER2.RTF';

PROC LOGISTIC DATA=CENTER_T2 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CENTER:FILLED ALL VARS: TIME 2';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
```

```
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                                /SELECTION=STEPWISE
                                                 CTABLE
                                                 RSQUARE
                                                 LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                 OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                 ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 3*/
ODS RTF FILE='CENTER3.RTF';

PROC LOGISTIC DATA=CENTER_T3 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CENTER:FILLED ALL VARS:TIME 3';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                                /SELECTION=STEPWISE
                                                 CTABLE
                                                 RSQUARE
                                                 LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                 OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
```

```
                                                    ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 4*/
ODS RTF FILE='CENTER4.RTF';

PROC LOGISTIC DATA=CENTER_T4 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR CENTER:FILLED ALL VARS:TIME 4';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                                   /SELECTION=STEPWISE
                                                    CTABLE
                                                    RSQUARE
                                                    LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                    OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                                    ;
RUN;

ODS RTF CLOSE;

/*RUNNING ANALYSIS ON RIGHT FIELD FOR EACH TIME FRAME*/

/*TIME FRAME NUMBER 1*/
ODS RTF FILE='RIGHT1.RTF';
```

```
PROC LOGISTIC DATA=RIGHT_T1 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR RIGHT:FILLED:ALL VARS';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                          /SELECTION=STEPWISE
                                           CTABLE
                                           RSQUARE
                                           LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                           OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                           ;
RUN;

ODS RTF CLOSE;

/*RIGHT T1 DOESNT HAVE AN OUTPUT BECAUSE ALL PLAYERS WERE ADMITTED INTO
THE HALL OF FAME.
THER EIS NO WAY TO RUN A LOGISTIC REGRESSION*/
/*TIME FRAME NUMBER 2*/
ODS RTF FILE='RIGHT2.RTF';

PROC LOGISTIC DATA=RIGHT_T2 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR RIGHT:FILLED ALL VARS: TIME 2';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
```

```
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                         OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL
                                         ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 3*/
ODS RTF FILE='RIGHT3.RTF';

PROC LOGISTIC DATA=RIGHT_T3 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR RIGHT:FILLED ALL VARS:TIME 3';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                        /SELECTION=STEPWISE
                                         CTABLE
                                         RSQUARE
                                         LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
```

```
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL

                                                ;
RUN;

ODS RTF CLOSE;
/*TIME FRAME NUMBER 4*/
ODS RTF FILE='RIGHT4.RTF';

PROC LOGISTIC DATA=RIGHT_T4 OUTEST=EST_P COVOUT;
TITLE'LOGISTIC REGRESSION FOR RIGHT:FILLED ALL VARS:TIME 4';
MODEL HOF(EVENT='1')= YRS GAMES_PLAYED PLATE_APPEARANCES AT_BATS RUNS
HITS DOUBLES TRIPLES HOME_RUNS RBIS STOLEN_BASES CAUGHT_STEALING
WALKS STRIKE_OUTS BATTING_AVERAGE ON_BASE__ SLUGGING__
OPS OPS_ TOTAL_BASES GDP HIT_BY_PITCH SACRIFICE_HITS SACRIFICE_FLY
INTENTIONAL_WALKS
G GS CG Inn Ch PO A E DP Fld_ RTOT RDRS RTOT_YR RDRS_YR RF_9 RF_G
lgFld_ lgRF9 lgRFG

                                               /SELECTION=STEPWISE
                                                CTABLE
                                                RSQUARE
                                                LACKFIT;/*GOODNESS OF FIT
TEST FOR CHI SQUARED ANALYSIS*/
                                                OUTPUT OUT=PRED P=PHAT
LOWER=LCL UPPER=UCL

                                                ;
RUN;

ODS RTF CLOSE;
```