2011

Individual Batting Factors Influencing a Team's Probability of Winning

David Matthew Rygh

INDIVIDUAL BATTING FACTORS INFLUENCING A TEAM'S PROBABILITY OF WINNING

A Thesis or Project

Submitted

in Partial Fulfillment

of the Requirements for the Designation

University Honors

David Matthew Rygh

University of Northern Iowa

May 2011

This Study by: David Rygh

Entitled: Individual Batting Factors Influencing a Team's Probability of Winning

has been approved as meeting the thesis or project requirement for the Designation

University Honors

5/2/11
Date

Mark Ecker, Honors Thesis/Project Advisor

5/6/11
Date

Jessica Moon, Director, University Honors Program

Introduction

Current Major League Baseball contracts are increasing rapidly in value. The factors that increase pay are primarily determined by the team owners' desire to win and players' performance. Understanding how an individual player contributes to winning could lead to owners signing players to more appropriate contracts. A statistical analysis, including a logistic regression analysis, was performed. The regression attempts to find how individual batting factors influence winning. The resulting model can be used to predict wins.

Literature Review

Four of the fourteen largest per-year salaries for Major League Baseball (MLB) players are new contracts beginning in 2011 (Cot's, n.d.). The salaries have recently skyrocketed, with the largest salaries approaching 30 million dollars per year. While these contracts seem exorbitant, they continue to grow. There are various causes for increase in salaries. Much of this can be attributed to higher revenues for the owners of the baseball teams, which then gets dispersed down to the players. However, it is also caused by the owners' desires to have the winning team. Many of these owners, especially George Steinbrenner, would do anything to win. Steinbrenner once stated that "winning is the most important thing in my life, after breathing" (Sports Illustrated, 2010). To accomplish this, bringing in the best players, regardless of cost, was deemed necessary. For large market teams, this could be justified as the revenues recoup these high costs. Meanwhile, small market teams are not able to compete salary-wise with these large market teams.

With the revenue and salary differences in Major League Baseball, a lower competitive balance is achieved. By paying significantly higher salaries to some of the top players, the large market teams are able to piece together a team that consistently wins. This lack of competitive balance has been studied to determine whether it is beneficial to team revenues, with some mixed results. Krautmann and Hadley (2006) found that "fans respond negatively when the game is dominated by perennial winners" (p. #). They found that, while statistically significant, this domination causes approximately 1% less gate receipts. The one concern with their findings is that they did not include aspects such as radio and television revenues.

Even if the team revenues are not as high due to the lack of competitive balance, some teams will continue to spend a lot to get the best players. Steinbrenner said this best: "What happens is that all your life you operated businesses in such a way that you could afford to buy a baseball team, and then you buy the team and ignore all of the business practices that enabled you to buy it" (Einwolf, 2004, p. 127). Because of this, there are often times contract disputes during the negotiations between players and teams. One such example was the situation with Albert Pujols going into the 2011 season. He was one of the top players in baseball history, hitting for both average and power. He was set to become a free agent at the end of the year. Prior to the 2011 season, he was in negotiations with the St. Louis Cardinals but failed to reach a new contract. The reasons for this are quite obvious: he feels he could make more money elsewhere and would command a higher salary. Teams such as the New York Yankees and Chicago Cubs would love to pay him lots of money so they may win more games. The idea is that a player as talented as Albert Pujols contributes a lot to a team producing wins.

While the situation of rising salaries runs across the whole league, the most intriguing aspect is with the top players. With them being paid so highly, teams should be curious as to how many wins they bring to the team. Very little research has been done on this exact field, especially with respect to a game-by-game analysis.

There has been research on areas involving the determination of who the best players were. For instance, David Berri (1999) performed an analysis on the National Basketball Association to determine the number of wins produced by each player. Berri used regression analysis to find marginal values of statistics that help and hurt the team. For all players, a predicted per minute production was calculated. Once this was calculated, per minute tempo and defensive factors were added. The positional average was then subtracted to get the per minute production relative to the average player at that position. The total number of games won throughout the season, 1189, was divided by total minutes played to get the average wins per minute. Average wins per minute was added to the per minute production, with the sum being multiplied by minutes played. The resulting number was the wins produced by the individual players. Berri (1999) checked to see how accurate the results were by summing all players on the team to determine the deviation from the number of wins by the team. The results gave fairly accurate predictions, but did have two errors of over seven games when compared to actual number of team wins. The interesting aspect of this paper was its finding of who contributed the most wins for their team: Dennis Rodman. Rodman was solely an excellent rebounder who brought few other skills to the team. While he certainly added a lot to the team, it was profound that he would add more wins than players like Michael Jordan and Karl Malone. Berri (1999) listed the top ten players for the regular season, which upon further research appears to have a bias. The top five were all power forwards or centers that average at least 10 rebounds per game. The next five did include three guards along with two power forwards, but these three guards rebounded very well for their positions.

Another study compared different players' years to determine which was better. To do this, Timothy Anderson and Gunter Sharp (2004) used data envelopment analysis (DEA) to find the best baseball player. DEA allows for viewing how efficient the top players are in comparison to other players in that particular year. The DEA finds an efficiency score, with being under one considered as super-efficient. Envelopment multipliers were also calculated. Any non-zero envelopment multiplier indicated what player or players he compares most to. Through the use of efficiency scores, direct comparisons can be made between two separate players in two separate years. Anderson and Sharp (2004) found an efficiency score of .31599 for Babe Ruth in 1920 and .44870 in 1921; Barry Bonds had an efficiency score of .82611 in 2001. The efficiency scores indicate that other players could only accomplish 32% or 45% of Babe Ruth's production, while other players could accomplish 83% of Barry Bonds' production. Anderson and Sharp (2004) warned about direct comparison, though, as the wording of the comparison would change the answer. If the comparison was an attempt to see what batter dominated the largest percentage of batters, the answer would change to who was compared against more. Domination of other players indicated that the player beat them in all areas of production. In these cases, neither Ruth nor Bonds would be the answer. Whichever player was compared to by the largest number of players would be the player that dominated the largest number of players.

The data envelopment analysis used by Anderson and Sharp (2004) used six statistics in total. The statistics it used were walks, singles, doubles, triples, home runs, and plate appearances. While this is adequate for determining what hitter is "best", more is needed in order to translate these statistics into wins. On top of these statistics, other aspects of the game must be accounted for. These aspects include pitching, defense, and baserunning. In order to effectively model wins, variables for these three should be included. Herman Demmink (2009) studied the impact of stolen bases on wins. He ran a regression involving just offensive statistics, steal attempts, and caught stealing. The regression used team statistics for the season and found that steal attempts did have a slight, but significant, positive effect on wins. Demmink's (2009) regression did not result in a particularly good model because of his inclusion of only offensive statistics. Demmink (2009) claimed that the omitted variables are not correlated with the offensive statistics, so the coefficients for offensive statistics would not change. One potential issue with his thought on this is that stolen bases have the potential of being correlated with defensive plays made. This correlation would be the result of one common factor improving both: speed.

Model

As the impact of the top players in the game is the subset I am interested in, analysis included only top players. Another aspect that was to be examined is the impact of different players. From this, I selected two individual premier players, as well as their best individual years. The two selected were Barry Bonds' 2001 season and Albert Pujols' 2009 season. Rather than using full season statistics like Berri (1999) and

Anderson and Sharp (2004) did, game statistics were used in this analysis. The use of game statistics is due to the desired results of what batting factors influence a team's probability of winning. In order to accurately model wins, statistics besides the individual player statistics must be considered. Because baseball is a team sport, there is more to winning than one player's performance. Two team variables were included: team errors and opposing pitcher earned run average (ERA). The number of team errors is a good indicator of how well the team played defensively. Opposing pitcher ERA is a good barometer of the quality of pitcher that is being faced. While other players may perform other offensive production, no covariate for this has been included. Runs batted in would be somewhat indicative of other players' performance. This could cause multicollinearity issues if total team hits were included.

Hypotheses

While there has not been much similar research, general knowledge about baseball suggests direction of correlations between the independent variables and the probability of winning. Any variable that increases the expected runs scored should be positively correlated, with the opposite being negatively correlated. Also, any variable that would increase the amount of runs given up would be negatively correlated. The predicted relationships between the dependent variable and the independent variables are shown in Table 1.

Table 1. Predicted Relationships

Independent Variable	Relationship Type
Plate Appearances	Unknown
At Bats	Unknown
Runs	Positive
Hits	Positive
Doubles	Positive
Triples	Positive
Home Runs	Positive
Runs Batted In	Positive
Walks	Unknown
Strikeouts	Negative
Hit By Pitch	Unknown
Stolen Bases	Positive
Team Errors	Negative
Opposing Pitcher ERA	Positive
Player	Unknown

Plate appearances and at bats may or may not be correlated with the probability of winning. Both of these would increase with the team scoring more runs, leaving more men on base, or by going into extra innings. Because of the extra innings impact, it is likely to not be highly correlated, but could potentially be positively correlated. Runs and runs batted in should be positively correlated with the probability of winning because scoring more runs increases the probability of winning. Hits, doubles, triples, home runs, and stolen bases should be positively correlated because they increase the likelihood of scoring runs. Because both Barry Bonds and Albert Pujols are excellent hitters, the impact of walks and hit by pitch is ambiguous. The ambiguity is because it puts them on base, which should increase the expected number of runs scored. However, it does take the bat out of their hands and relies on the hitters behind them to produce. The team errors variable is predicted to be negatively correlated because committing errors often results in more runs given up. The opposing starting pitcher's earned run average is expected to be positively correlated because a higher ERA is an indication of a poorer quality pitcher. A worse pitcher will likely give up more runs, and therefore allow for the opponent to have a higher chance of winning. The variable for player may or may not be correlated with the probability of winning. It was included to test whether the same production by different players had a different impact on the probability of winning.

Methodology

Initially, data was collected for the two players' seasons. The game logs were used from Baseball-Reference.com to collect over twenty game statistics that could be used in the analysis. The covariates that were collected and used as independent variables include plate appearances, at bats, runs, hits, doubles, triples, home runs, runs batted in, walks, strikeouts, hit by pitch, stolen bases, team errors, and the opposing pitcher's earned run average. A binary variable for which player the game was played by was included, with Barry Bonds having the value 0 and Albert Pujols having the value 1. These variables were used in order to determine what impact each player had on their respective teams' probability of winning the game. The variable for a win was a dummy variable, taking the value 0 for a loss and 1 for a win. Summary statistics were calculated for the dataset.

Rather than using all assumed variables, model-building techniques were used to determine the best models for a logistic regression. These techniques are used to eliminate insignificant independent variables. These methods include forward, backward, and stepwise regressions along with including the full model. Models were set up for three different cases: one for Barry Bonds, one for Albert Pujols, and one that included both players. With the model that included both players, a reduction in deviance test was run to determine whether the two players had a different impact on their team's probability of winning. Logistic regression was used for the analysis due to the dependent variable being binary. A logistic regression is bounded between 0 and 1 and gives a probability of outcome when used for prediction. Statistical Analysis Software (SAS) was used for the above analyses. The code used for SAS with the model for both players is included in Appendix A.

After the logistic regression was run, a cluster analysis was performed. The cluster analysis was done for two different scenarios: the players' best games and the teams' worst losses. The last analysis that was done was checking for autocorrelation. The cluster analysis and autocorrelation were performed in S-Plus.

Results

Summary Statistics

The raw data for the two players' seasons were collected from Baseball-Reference.com. Summary statistics were calculated for the variables. Table 2 shows the summary statistics with both players' games considered.

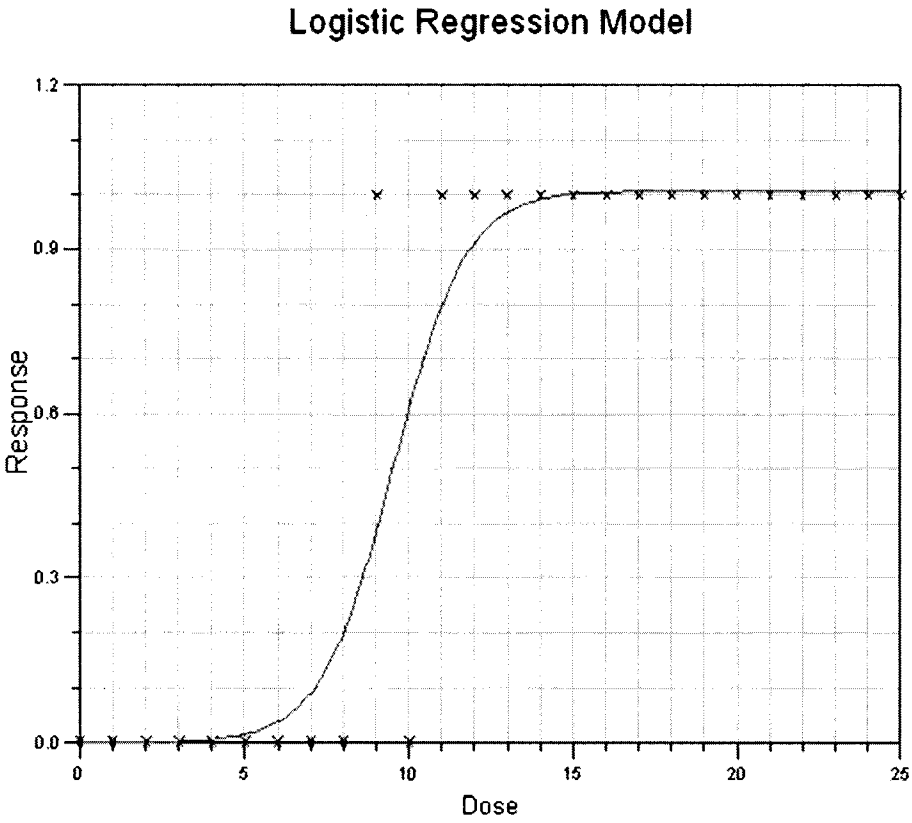
Table 2. Summary Statistics

Variable	Mean	Std Dev	Median	Min	Max
PA	4.3578275	0.8621740	4	1	8
AB	3.3354633	1.0374036	3	0	7
Runs	0.8083067	0.8595213	1	0	4
Hits	1.0926518	0.9204115	1	0	4
Doubles	0.2460064	0.4872000	0	0	2
Triples	0.0095847	0.0975870	0	0	1
Home Runs	0.3833866	0.6253041	0	0	3
RBIs	0.8690096	1.1761920	0	0	7
Walks	0.9329073	0.9191535	1	0	4
Strikeouts	0.5015974	0.6846513	0	0	3
HBP	0.0575080	0.2331835	0	0	1
Steals	0.0926518	0.2904081	0	0	1
Errors	0.6613419	0.8088101	0	0	4

Regression

A logistic regression analysis was used to determine which variables impacted the dependent variable. The logistic regression bounds the output between 0 and 1. The output for this regression is the probability that the team wins, given the certain independent variables. The output of the predicted variable is sigmoidal shape, as displayed in Figure 1. Figure 1 shows an example of the shape along with the observed, each displayed by an x.

Figure 1. Sigmoidal Shape



Source: DTREG

Before the regression was run, model-building techniques were used to determine which models fit best. The techniques used include the forward, backward, and stepwise selections.

The forward selection starts out with zero variables included and then iteratively adds in one variable at a time if it is significant. It continues with this process until no more variables that could be added are found to be significant based on p-values. A cutoff value of .15 was used. A cutoff value of .15 indicates that all of the significant variables' test statistics could have up to a 15% chance of occurring by pure chance alone. For model selection, a cutoff of .15 is typical.

The backward selection is very similar to the forward selection. It, however, starts with the full model and then eliminates insignificant variables. A cutoff value of .15 was used for the backward selection as well.

The stepwise selection is like a combination of the forward and backward selections. It starts out with zero variables. It then adds in the most significant variable if it has a p-value under .15. All variables in the model get checked after another one gets added in to ensure it is still significant. If it is insignificant, it is removed, and the selection continues.

Each of these three model-selection techniques was used in three scenarios. The first was the combination of the seasons by Albert Pujols and Barry Bonds. The second was Albert Pujols' season; the third was Barry Bonds' season. Through the use of these three regressions, it was possible to determine, in multiple ways, whether each player had a different impact on their team winning. First, a reduction in deviance test was performed to determine if the binary variable for player is significant. Deviance is a measure of how well the model fits the actual data. A reduction in deviance indicates that the model is better fitting. With the reduction in deviance test performed for the player variable, there was 0 reduction in deviance. The critical value of 5.024 was greater than the test statistic, indicating that the variable is not significant. The player variable not being significant indicates that there is no difference in how each variable for each player impacts the team winning. The significant variables decided by each method for each situation are displayed in the Table 3. The resulting models are displayed in Table 4.

Table 3. Significant Variables

Player	Forward	Backward	Stepwise
Both	Runs, Home Runs, RBIs, Errors	Runs, Home Runs, RBIs, Errors	Runs, Home Runs, RBIs, Errors
Albert Pujols	Runs, RBIs, Errors, Walks	Runs, RBIs, Errors, Walks	Runs, RBIs, Errors, Walks
Barry Bonds	Runs, Home Runs, Errors	Runs, Home Runs, Errors	Runs, Home Runs, Errors

Table 4. Models

Both	-0.4321 + 1.4361*Runs – 1.0547*Home Runs + 0.4301*RBIs – 0.5640*Errors
Albert Pujols	-0.8870 + 1.1158*Runs + 0.5201*RBIs + 0.4524*Walks – 0.4659*Errors
Barry Bonds	-0.1447 + 1.4923*Runs – 0.8343*Home Runs – 0.6127*Errors

With the logistic regression, the value given by the equations above can be transformed into the probability of winning the game. The function used for this is:

$$f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}}$$

In this function, z is the value from the above models. When z is 0, the function equals 0.5. This indicates that there is a 50 percent chance that the team won the game, given the certain values of the explanatory variables. As z increases, the function approaches 1. As z decreases, the function approaches 0.

Next, the models for Pujols and Bonds were directly compared. While both of the models included runs and errors as explanatory variables, they were still quite different overall. For example, when a solo home run is hit, the player scores one run and gets one run batted in. In the Pujols model, the effect on z is an increase of 1.6359. In the Bonds model, the effect is an increase of 0.658. The impact of an increase in each variable for each model is shown below. These results are multiplying factors for the odds. The factor for home run is indicative of a solo home run. The factor would be higher if others were on base. The results are shown in Table 5.

Table 5. Odds Ratios

	Both	Albert	Barry
Run	4.2	3.05	4.45
Home Run*	2.25	5.13	1.93
RBI	1.54	1.68	N/A
Error	0.57	0.63	0.54
Walks	N/A	1.57	N/A

This is not the biggest difference though, as the variables in the model are different. For Pujols, the team has a higher probability of winning as he is walked. For Bonds this is not the case.

The final model that considered both players has a few interesting aspects. One of these is the lack of the player variable. The variables absence indicates that for both players, the variables have the same impact on win probability. The negative regression coefficient for home runs is contrary to what was expected before the analysis was performed. The coefficient can easily be explained, however. As explained above, home runs are directly correlated with runs and runs batted in. This is the result of a home run also counting as a run and having at least one run batted in. In reality, the variables should be broken down to runs not scored from home runs and non-self driven in runs to accurately show the relationship. In that scenario, the coefficients would adjust, but result in the same exact ending probability of winning. Table 6 shows the possible covariates, the expected relationship, and the relationship found.

Table 6. Significant Relationships

Independent Variable	Predicted Relationship Type	Relationship Type Found
Plate Appearances	Unknown	None
At Bats	Unknown	None
Runs	Positive	Positive
Hits	Positive	None
Doubles	Positive	None
Triples	Positive	None
Home Runs	Positive	Negative*
Runs Batted In	Positive	Positive
Walks	Unknown	None
Strikeouts	Negative	None
Hit By Pitch	Unknown	None
Stolen Bases	Positive	None
Team Errors	Negative	Negative
Opposing Pitcher ERA	Positive	None
Player	Unknown	None

As can be seen by Table 6, the resulting model follows as was predicted. Many of the variables were found to be insignificant, but the direction of the significant was in the direction that was expected. The home run variable by itself was negative, but must be considered with at least run batted in and one run. Taking that into consideration, the actual result is positive, even though the coefficient is negative.

Cluster Analysis

Following the logistic regression analysis, a cluster analysis was performed. A cluster analysis shows which observations are "clustered" based on their similarity. For the cluster analysis, two situations were looked at. The first situation was each player's ten best hitting games. The ten games were decided by home runs in the game. For both players, the games were all multi home run games. The second situation was each teams worst losses. For the worst losses, games lost by six or more runs were included. The dendogram for the best games displayed clusters based on wins and losses. The dendogram for worst losses displayed clusters based on player.

Autocorrelation

In an attempt to see if either of the players had hot streaks, autocorrelation was checked. Autocorrelation is the correlation between a single variable at various lags in games played. Checking autocorrelation allows for determination of if patterns of good games following good games or bad games following bad games exist. The autocorrelation function was used to check for these streaks in home runs, hits, strikeouts, and total bases. For both Albert Pujols and Barry Bonds, the autocorrelation found almost no relationship whatsoever between games. For all of the tests, the autocorrelation was found to be insignificant with a few exceptions. The exceptions include a few cases in which the autocorrelation after ten or more games was found to be significant. There is no basis for such a finding, and it is deemed a coincidence. An example of the output of the autocorrelation is in Appendix B.

Prediction

To determine the effectiveness of this analysis, the models were used to predict wins. For both players, the probability of winning was calculated for both the individual model and the model including both players. These probabilities were then considered to be a win if they were greater than or equal to 0.50. The resulting predicted wins were compared to actual wins. Table 7 shows statistics on the predictions, along with actual wins.

Table 7. Predictions

Player	Model	Actual Wins	Predicted Wins	Correct Predictions
Albert Pujols	Both	90	84	114
	Individual	90	94	116
Barry Bonds	Both	83	78	104
	Individual	83	67	103

Considering that a hard cutoff was used, these predictions are very reasonable. This is because a probability of 50% is just as likely to be a win or loss, but must be assigned one value or another. Assessing just the observations in which the predicted probability was far enough away from 50% was a possibility, but would have thrown out data. Because of this, the hard cutoff was the method used for prediction.

Conclusion

This statistical analysis found four variables that were significant in determining a baseball team's probability of winning. The variables found include runs, home runs, runs batted in, and team errors. Runs and runs batted in both were found to have a positive relationship while home runs and team errors were negative. The variable for home runs did not take into account runs and runs batted in, so its actual effect is not perverse. This analysis also found that Albert Pujols and Barry Bonds did not have a different impact on their teams' probability of winning when they perform the same.

References

Anderson, T., & Sharp, G. (1997, January). A new measure of baseball batters using DEA. *Annals of Operations Research, 73*, 141-155.

Baseball-Reference. (n.d.). Major League Baseball Statistics and History. *Baseball-Reference.com*. Retrieved from http://www.baseball-reference.com/

Berri, D. (1999, December). Who is 'most valuable'? Measuring the player's production of wins in the National Basketball Association. *Managerial and Decision Economis, 20*, 411-427.

Cot's Baseball Contracts. (n.d.). *Cot's baseball contracts*. Retrieved from http://mlbcontracts.blogspot.com/2000/05/most-lucrative-contracts.html

Demmink, H. (2010, March). Value of stealing bases in Major League Baseball. *Public Choice, 142*, 497-505.

Einwolf, K. (2004, May). Is winning everything?: A data envelopment analysis of Major League Baseball and the National Football League. *Journal of Sports Economics, 5*, 127-151.

Krautmann, A., & Hadley, L. (2006, June). Dynasties versus pennant races: competitive balance in Major League Baseball. *Managerial and Decision Economics, 27*, 287-292.

[Logistic regression graph]. Retrieved April 29, 2011, from: http://www.dtreg.com/doseresponse.jpg

Sports Illustrated. (2010, July). George Steinbrenner's most colorful quotes. *Sports Illustrated.* Retrieved from http://sportsillustrated.cnn.com/2010/baseball/mlb/07/13/steinbrenner.quotes/index.html

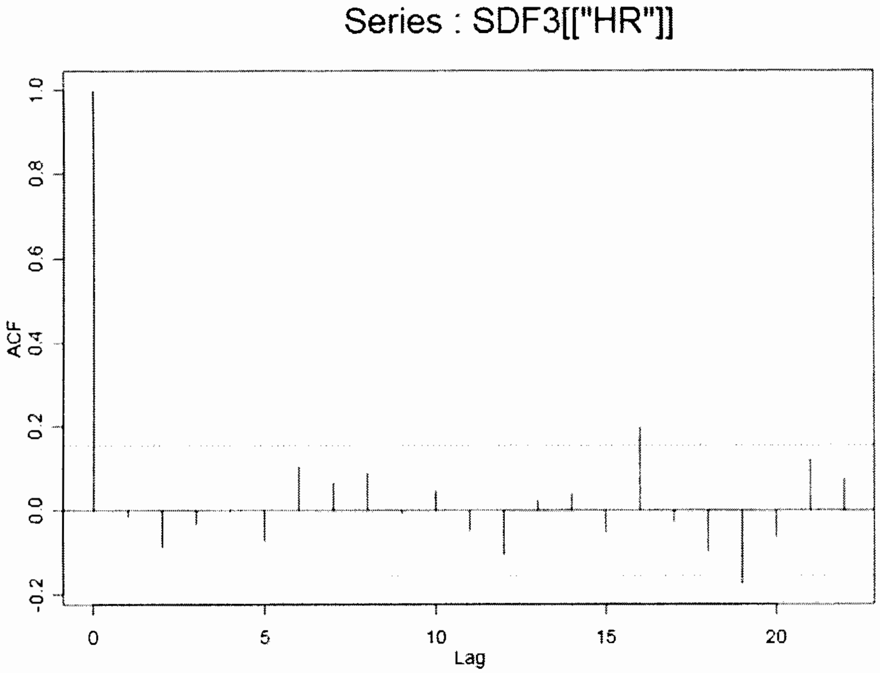
Appendices

Appendix A: SAS Code

```
data baseball;
input win pa ab runs hits double triple hr rbi bb so hbp sb ba obp
error era player;
cards;

<DATA>


proc sort data=baseball;
by descending win;

proc logistic order=data;
model win = pa ab runs hits double triple hr rbi bb so hbp sb ba obp
error era player / selection = forward slentry = .15;

proc logistic order=data;
model win = pa ab runs hits double triple hr rbi bb so hbp sb ba obp
error era player / selection = backward slstay = .15;

proc logistic order=data;
model win = pa ab runs hits double triple hr rbi bb so hbp sb ba obp
error era player / selection = stepwise slentry = .15 slstay = .15;
run;
```

Appendix B: ACF Output

```
       *** Autocorrelations for series SDF4[["H"]] ***

Call: acf(x = SDF3[["H"]], type = "correlation", plot = FALSE)

Autocorrelation matrix:
   lag    SDF4
 1   0  1.0000
 2   1  0.1258
 3   2 -0.1265
 4   3  0.0417
 5   4  0.1004
 6   5 -0.0427
 7   6 -0.0849
 8   7  0.0074
 9   8 -0.0849
10   9 -0.0846
11  10 -0.0342
12  11 -0.0596
13  12 -0.0848
14  13 -0.0258
15  14 -0.0846
16  15  0.0161
17  16  0.1674
18  17  0.0749
19  18 -0.0681
20  19  0.0501
21  20  0.1087
22  21  0.0502
```


Series : SDF3[["HR"]]
ACF
Lag