What Makes a Winning Baseball Team and What Makes a Playoff Team?

Javier Lopez New Mexico State University
Daniel J. Mundfrom Eastern Kentucky University
Jay R. Schaffer University of Northern Colorado

Team statistics from all 30 teams in Major League Baseball were analyzed to determine what makes a winning baseball team and what makes a playoff team. Thirty-two statistics in all, including batting, fielding, and pitching statistics, were used in a multiple linear regression and discriminant analyses. The regression procedure was to determine what makes a winning team, while the discriminant analyses were used to see what makes a playoff team. On-base percentage plus slugging (OPS) and earned run average (ERA) were fit on wins in the regression model with an $R^2 = 0.83$. The discriminant analyses distinguished different statistics in the National and American Leagues for discriminating between playoff and non-playoff teams. ERA and OPS were discriminating factors in the National League, while saves, on-base percentage, and earned run average were factors in the American League.

Every year, 30 Major League Baseball (MLB) teams strive to make the playoffs and World Series. Sixteen teams are from the National League and 14 are from the American League. Each league is split into three divisions, the East, Central, and West. In order to make the playoffs, a team must either win their division or win the wild card spot. That is, at the end of the 162 regular season games, they must either have the most wins in their division or have the most wins among the teams that did not win their division. Essentially, in order to be a successful team in MLB, you must win. General Managers (GM) in MLB can be fired because they are unable to build a winning baseball team. The question that every GM in MLB should be asking himself as they build their team each year is, "*What makes a winning baseball team*?"

Talsma (1999) answers this question in a simple and forthright manner. In his study, he determined that run differential (RDIFF = runs scored – runs allowed) is the most important statistic in determining how many wins a baseball team will have. The higher the run differential, the more wins a team will have. But if we are trying to build an MLB team, we cannot directly sign "runs" to our team. That is, we can only sign players. For instance, if we want to sign a free agent such as Manny Ramirez, there are many statistics that we can look at to determine his worth. However, the question may well be "which statistics are the ones that appear to contribute to winning and runs scored?" Should a GM sign him, and if so, on what should he base his decision if he cannot determine how many runs he is worth? In this study, we take 32 team statistics into account and attempt to find which ones best determine the wins a team will have in a given year. We do this by using past research to help determine which statistics should be used to fit a multiple linear regression model for all data collected between 1995 - 2009.

The second objective of this study is to determine which statistics are best for predicting which teams will make the playoffs and which teams will not. There have been some models to predict divisional winners. Barry and Hartigan (1993) used Markov chain sampling to determine future strengths of teams and future outcomes of games to predict divisional winners. This research was done before MLB split into 3 divisions and the addition of a wild card team in the playoffs. Barry and Hartigan only had a total of 4 playoff teams each year when this research was done, as opposed to the present day when we now have 8 playoff teams each year. For this reason we only use data after 1994 when the new wild card rule took effect. In this study, we use discriminant analysis to find out which statistics discriminate between teams that made and missed the playoffs.

Review of Literature

Talsma (1999) used a simple linear regression analysis to determine what made a winning baseball team. He initially explored the relationships between wins and hits, homeruns, runs, and opponents' runs. All four of these statistics are offensive statistics. After finding that the relationship between those four variables and winning was weak, he decided that maybe offense was not the only way a team could win. He found that run differential (RDIFF = runs scored – runs allowed) is the most important statistic in determining how many wins a baseball team will have. The higher the RDIFF, the more wins a team will have. His regression equation was:

$$\widehat{Wins} = 0.088 * RDIFF + 80.964$$

Talsma’s results are vital information for use in this study. If run differential is the best predictor for wins, we would then need to find out which player and team statistics best predict runs scored and runs allowed.

There is a plethora of statistics that we can examine in baseball to evaluate performance. Batting average (BA), on-base percentage (OBP), home runs (HR), earned run average (ERA), strike outs (SO), on-base percentage plus slugging(OPS), and fielding percentage (Fld%) are just a few examples. But which ones are the most valuable?

A study by Cover and Keilers (1977) used a batter’s cumulative statistics to determine which batters were the best of all time. They used at-bats, walks, singles, doubles, triples, homeruns, and outs to determine what they called an offensive earned run average (OERA). Cover and Keilers defined OERA as the number of earned runs per game that a player would score if he batted in all nine positions in the line-up. After getting the OERA for all players, they then could be compared to each other to determine which players were actually the best batters in the league. The higher the OERA, which means they would score more runs, the better the batter.

Another study by Koop (2002) used an output aggregator to compare players. This method weighed multiple batting statistics to output a value between 0 and 1, where 1 would be the best player and 0 the worst. These two methods of determining a player’s performance were very different, but the consensus between these studies is that there were multiple statistics that contributed to the performance of a player.

There has been little published in regard to what statistics best predict playoff teams. As stated earlier, Barry and Hartigan (1993) used Markov chain sampling to determine future strengths of teams and future outcomes of games to predict divisional winners. They used information based on strength of teams to determine divisional winners. Our objective in the current study is to determine which baseball statistics may be able to be used to predict which teams will make the playoffs and which teams will not.

Table 1. Offensive and Defensive Statistics.

Offensive Statistics		
Abbreviation	Statistic	Mathematical Definition
PA	Plate Appearances	
AB	At-Bats	
H	Offensive Hits	
2B	Doubles Hit	
3B	Triples Hit	
HR	Home Runs Hit	
SB	Stolen Bases	
CS	Caught Stealing	
BB	Bases on Balls	
SO	Strike Outs	
BA	Batting Average	$\frac{\text{Hits}}{\text{AB}}$
OBP	On Base Percentage	$\frac{\text{(H+BB+HBP)}}{\text{(At-Bats+BB+HBP+SF)}}$
SLG	Slugging Percentage	$\frac{\text{(1B+2*2B+3*3B+4*HR)}}{\text{AB}}$
OPS	On Base Plus Slugging	OBP+SLG
E	Errors Committed	
DP	Double Plays Turned	
BatAge	Batter's Age	
Defensive Statistics		
Fld%	Fielding Percentage	$\frac{\text{(Putouts+Assists)}}{\text{(Putouts+Assists+Errors)}}$
ERA	Earned Run Average`	$\frac{\text{(Earned Runs*9)}}{\text{IP}}$
CG	Complete Games	
SHO	Shutouts	
SV	Saves	
IP	Innings Pitched	
H	Hits Allowed	
ER	Earned Runs Allowed	
HR	Home Runs Allowed	
BB	Bases on Balls Allowed	
SO	Strike Outs	
WHIP	Walks & Hits Per Innings Pitched	$\frac{\text{(BB+H)}}{\text{IP}}$
SO/9	Strike Outs Per 9 Innings	$\frac{\text{(9*Strikeouts)}}{\text{IP}}$
HR/9	Home Runs Per 9 Innings	$\frac{\text{(9*HR)}}{\text{IP}}$
PitchAge	Pitchers Age	

Methods

The following 32 statistics from Table 1 were collected for all teams' from the 1995 - 2009 seasons (Baseball-Reference.com, 2010). The data used in these analyses were the final season-ending values of each of these statistics for each team collectively for each of these 15 seasons.

Data were analyzed using SAS version 9.1.3 software using PROC GLM, PROC REG, PROC DISCRIM, and PROC STEPDISC. Data were initially explored graphically and, subsequent to model fitting, residual analysis was conducted. All 32 statistics for both American and National League teams were initially used in the multiple linear regression analysis. After eliminating variables to remove multicollinearity from the data, variables with large p-values were taken out of the model one at a time using the method described below. The statistical significance level for the regression was defined for $p < 0.05$.

Beginning with the full model that contained all 32 variables, an examination of the pairwise correlations and variance inflation factors (VIF) identified several pairs or groups of variables that were collinear (i.e., pairwise correlations > .70 and VIFs > 10). By removing variables one at a time and re-running the model with one fewer variable at each step, the multicollinearity was successfully removed with the elimination of 12 variables, leaving 20 predictor variables in the model with only one having a VIF slightly larger than 10 ($R^2 = .889$, adjusted $R^2 = .885$). The remaining predictors were plate appearances (PA), doubles (DB), triples (TR), stolen bases (SB), caught stealing (CS), offensive strikeouts (OSO), batting average (BA), on base plus slugging (OPS), errors (E), double plays, (DP), batter's age (BatAge), earned run average (ERA), complete games (CG), shutouts (SHO), saves (SV), hits (H), home runs (HR), walks (BB), defensive strikeouts (SO), and pitcher's age (PitchAge). With this model, the unique contribution of each variable to the explanation of the variance in the number of wins was examined (using Type III Sums of Squares in SAS) and the variable that made the smallest unique contribution (i.e., provided its p-value was less than .05) was dropped from the model.

The remaining 20 variables were re-analyzed and the variable with the smallest, non-significant unique contribution to the model was dropped. This process was continued, at each stage dropping only the one variable that made the smallest, non-significant, contribution to the explanation of the variation in the number of wins, until only the variables that made significant unique contributions remained. The final model, a multiple linear regression of OPS and ERA on Wins was fit.

Stepwise discriminant analyses were performed on the 32 statistics to determine which statistics discriminate between teams that made and missed the playoffs for both the National League and American Leagues separately. Separate analyses were used because the American League uses a designated hitter as opposed to the National League, which does not. The discriminant procedure was run three times for each league. Each time the significance level (i.e., statistical significance level for entry and significance level for removal were set to be equal) was set to be different. The three different significance levels used were $p < 0.1$, $p < 0.05$, and $p < 0.01$. This process was used to determine whether more predictors would lower the total probability of misclassification (TPM). If the TPM stayed the same when there were fewer predictors, the extra predictors were deemed unnecessary because they were not contributing to better predicting classification. The jack-knife (cross-validation) method was performed in SAS to estimate the TPM. Both linear and quadratic models for the jack-knife method were evaluated to determine which one performed better based on the TPM. Lower TPM was deemed to be better than higher.

Results

A model to predict the number of wins based on OPS and ERA was considered. Because these two variables were the only ones that survived the variable-identification process, it could be surmised that these two statistics are in some way representing offensive prowess (i.e, OPS) and defensive prowess (i.e., ERA). For a team to have enough wins at the end of the season to make the playoffs, they must be good both offensively and defensively so it made sense that these two variables would be good predictors of the number of games a team wins during the season. Some teams may win a lot of games because they have very potent offenses, and others may win fewer games because they are less productive at the plate. Also, some teams may win a lot of games because they have good pitching and defense, and others may win less often because they do not. It is also possible that some teams that may be only "average" on both offense and defense may win a lot of games because they are able to score just a few more runs than they give up on enough occasions during the season so that they win a lot of games. Consequently, it seemed prudent to consider an interaction between these two variables as an addition to this model.

A model was subsequently fit that contained the two variables identified earlier, OPS and ERA, and their interaction to predict the number of wins. The results of this analysis showed that the interaction between OPS and ERA was not statistically significant (p = 0.8991. Consequently, it was dropped from the model. The final model from the multiple linear regression, containing only OPS and ERA as predictors, provided a good fit (F = 1044, p < 0.0001, R^2 = 0.83) and yielded the following estimated equation:

$$\widehat{Wins} = -10.45 + 213.62 * OPS - 15.98 * ERA$$

with standard errors for the intercept = 5.09, OPS = 6.5, and ERA = 0.45.

Results for the discriminant analysis varied with each league. Table 2 shows that the discriminant analysis did a credible job of classifying the American League teams properly. We see that 80% of the teams that made the playoffs were actually classified as making the playoffs and 89.12% of the teams that missed the playoffs were actually classified as missing the playoffs. Table 3 indicates that the discriminant analysis did not work quite as well in the National League as compared to the American League in terms of correctly classifying the teams that made the playoffs (66.3%). However, for the National League teams, 90.96% of the teams that missed the playoffs were classified correctly; just slightly more than in the American League.

Table 2. American League Classification Results.

American League	Classified as Type		
From Type	Made Playoffs	Missed Playoffs	Total
Made Playoffs	48	12	60
	80.00%	20.00%	100%
Missed Playoffs	16	131	147
	10.88%	89.12%	100%
Total	64	143	207
	30.92%	69.08%	100%

Table 3. National League Classification Results.

National League			
From Type	Made Playoffs	Missed Playoffs	Total
Made Playoffs	38	22	60
	63.33%	36.67%	100%
Missed Playoffs	16	161	177
	9.04%	90.96%	100%
Total	54	183	237
	22.78%	77.22%	100%

In Table 4, we see that for all cases, the linear function outperformed the quadratic. That is, the TPM was lower for the linear function in each case. The TPM stayed the same within league and model even when the stepwise significance level was decreased. The only change concerned the number of predictor variables identified as important discriminators.

Discussion

From our results, we found that statistics indeed can help us determine what makes a winning baseball team. OPS and ERA accounted for 82.57% of the variation in determining wins for a Major League baseball team. This result is consistent with what Talsma (1999) found in that both offense and defense are needed factors to succeed. Within our model, OPS is used as the offensive statistic and ERA is used as the defensive statistic. This fact may provide assistance to General Managers for potentially basing their decisions for personnel drafting and trading, as well as signing free agents. This is not to say that other statistics, such as stolen bases or strikeouts, do not contribute to winning; as all offense contributes to scoring runs and winning. However, when there are millions of dollars at stake, as there is in Major League Baseball, it is suggested by this analysis that General Manager's might benefit from looking at these two statistics before any others when making a decision to pursue a prospective player.

In the second set of analyses, we saw in both the National and American Leagues that even as the level of significance for selecting discriminating variables went down, the TPM remained the same. At p < 0.1, the TPM was the same at p < 0.01. What does change, is the number of discriminants, which even with this change, we can see that there is no change in the number of incorrectly classified teams. This result brought us to the conclusion that we could drop the extra variables because they were not contributing to the classification.

The next thing to notice in the results was that there were two different sets of statistics that discriminated between playoff teams and non-playoff teams. The National League followed our regression model; where the OPS and ERA were the best predictors in determining whether or not a team made the playoffs. However, the American League discriminant analysis indicated that saves (SV), on-base percentage (OBP), and earned run average (ERA) were the best predictors. The question then became, “Why were the results different between the two leagues?”

Let us first examine the difference between OBP and OPS. OBP is the percentage of times a batter gets on base by walking, hitting safely, or being hit by a pitch. OPS is equal to OBP plus slugging percentage (SLG%). Slugging percentage is the tell-tale statistic that takes into account a batter's power. A batter who hits mostly singles will have a lower SLG% than a batter who hits more doubles, triples, and home runs. These results indicate that OPS is more important in the National League and OBP is more important in the American League.

Table 4. Results from the Discriminant Analysis

Discriminant Analysis Cross-Validation Results				
League	**Model**	**TPM**	**Stepwise Sig.**	**Predictors**
NL	Linear	0.1595	p≤0.1	ERA,OPS,2B,Fld%
NL	Quad	0.2337	p≤0.1	ERA,OPS,2B,Fld%
NL	Linear	0.1595	p≤0.05	ERA, OPS, 2B
NL	Quad	0.2337	p≤0.05	ERA, OPS, 2B
NLl	Linear	0.1595	p≤0.01	ERA, OPS
NL	Quad	0.2337	p≤0.01	ERA, OPS
AL	Linear	0.1349	p≤0.1	SV,OBP,ERA,DP
AL	Quad	0.2676	p≤0.1	SV,OBP,ERA,DP
AL	Linear	0.1349	p≤0.05	SV,OBP,ERA
AL	Quad	0.2676	p≤0.05	SV,OBP,ERA
AL	Linear	0.1349	p≤0.01	SV,OBP,ERA
AL	Quad	0.2676	p≤0.01	SV,OBP,ERA

One reason there may be a difference is because the American League uses a designated hitter (DH) and the National League does not. That is, in the American League, the pitchers do not hit. The designated hitter is used in the lineup instead. Whereas, in the National League, the pitchers bat for themselves until a substitution is made that removes the pitcher from the game. Pitchers, in general, are the least proficient batters in the league. The use of the DH in the American League contributes to why most ERAs are lower in the National League. That is, National League pitchers face a lesser quality batter 1 out of every 9 batters in the lineup. It could be that it is for this reason OPS is more important in the National League.

Let us consider an example. We will assume that we have two outs and a man on first base. The 8th man in the lineup, who has a low slugging percentage, is up to bat and the pitcher is on deck to bat next. If the 8th man hits a single, the runner advances to 2nd base only. The next batter is the pitcher and in most cases, he is assumed to be an easy out. So if he strikes out or fails to hit safely, then the inning is over and 2 men are left on base. If, however, the 8th man in the order happens to be more of a power hitter and he hits a double, the runner may score a run. Then, once again, the pitcher, who is the next batter, strikes out or does not hit safely to end the inning. The difference being that a run scores with a double instead of a single. This scenario occurs daily in the National League.

In the American League, this same scenario does not happen as often because instead of having the pitcher bat, the league uses a DH who in most cases is a better batter. In the same hypothetical situation, if the 8th batter only hits a single, we would again have a man on first and second. Only this time, in the American League, the pitcher or "easy out" is not up to bat. Instead, we would have a higher quality batter up who even if he only singles, he drives in a run; whereas in the National League, the pitcher is much more likely to make an out. The OPS statistic takes into account a batter’s power whereas OBP does not. We can see from the previous simple example, why power is an important aspect and, therefore, OPS was more important in the National League in the results.

Our results also show that saves are a significant factor in the American League and not in the National League. This result could also be due to the quality of batters that a pitcher must face in the American League. However, our previous example does not quite apply to this situation. Saves are credited to a pitcher when:

1. He is the finishing pitcher in a game won by his team;
2. He is not the winning pitcher;
3. He is credited with at least ⅓ of an inning pitched; and
4. He satisfies one of the following conditions:
   a. He enters the game with a lead of no more than three runs and pitches for at least one inning.
   b. He enters the game, regardless of the count, with the potential tying run either on base, at bat, or on deck.
   c. He pitches for at least three innings.

In the National League, a pinch hitter (PH) can bat for a pitcher only if the pitcher is substituted out of the game. After the PH hits, a different pitcher must come into the game to pitch. When it next becomes the pitcher's turn to bat, he must bat, or another PH can hit for him, and he will be substituted out. Once a pitcher is taken out of the game, he cannot come back in. The same goes for the PH. Any one player can only pinch hit once in a game (i.e., unless his turn comes around again in the same inning or he enters the game as a position player after completing his pinch hit). In the late innings of a game, the pitchers in the National League are usually substituted for to try and maximize runs scored. By having a better batter pinch hit for the pitcher, you eliminate what is often an "easy out" in the lineup.

Most likely, pitchers that are in save situations, will not have an easy out, but the quality of hitter may still differ from the National League to the American League. The DH in the American League hits every time around in the lineup. A pinch hitter in the National League sits most of the game until a substitution is needed. When a substitution is made, the pinch hitter usually goes up to bat one time and then his job is over for the day. For this reason, it is probably reasonable to say that it is more difficult to be a pinch hitter than a designated hitter. We may also guess that it is easier to get most PHs out than it is to get DHs out. A pinch hitter has sat most of the game and only has one chance to hit in a game whereas the designated hitter has probably already hit three or four times before a closing pitcher comes in to attempt a save. Since it is more difficult to get a designated hitter out, it becomes more crucial to have a good closer to get a save in the American League than in the National League. Consequently, the use of the DH in the American League versus the PH in the National League could be contributing to why our results show that SVs are more important in the American League than in the National League.

Conclusion

In conclusion, we can see that certain statistics are vital in determining wins for a baseball team and determining which teams will and will not make the playoffs. We can predict wins based on OPS and ERA. We can also determine that OPS and ERA are significant factors for predicting if a team makes the playoffs in the National League, whereas Saves, OBP, and ERA are significant factors in predicting whether a team makes the playoffs in the American League. Although our discriminant analysis procedure did predict 80% of teams correctly to make the playoffs in the American League, it only predicted 63.33% correctly for the National League. This outcome is not bad and is certainly better than simply guessing or flipping a coin. Perhaps future research can identify other variables that can improve this percentage.

References

Barry, D., & Hartigan, J. (1993). Choice models for predicting divisional winners in Major League Baseball. *Journal of the American Statistical Association, 88*(423), 766-774.

Baseball-Reference.com. (2010). *Major League Baseball statistics and history*. Retrieved from http://www.baseball-reference.com.

Cover, T., & Keilers C. (1977). An offensive earned-run average of baseball. *Operations Research*, *25*(5), 729-740.

Koop, G. (2002). Comparing the performance of baseball players: A multiple-output approach. *Journal of the American Statistical Association, 97*(459), 710-721.

Talsma, G. (1999). Data analysis and baseball. *Mathematics Teacher, 92*(8), 738-742.

Send correspondence to: Javier Lopez
New Mexico State University
Email: javmlopez@yahoo.com