A Mathematical Model to Predict Award Winners

Author(s): Rebecca L. Sparks and David L. Abrahamson

Source: *Math Horizons*, April 2005, Vol. 12, No. 4 (April 2005), pp. 5-8, 13

Published by: Taylor & Francis, Ltd. on behalf of the Mathematical Association of America

Stable URL: https://www.jstor.org/stable/25678541

"Is it possible to look at the information for each candidate and combine it in such a way that we can predict how the voters will rank the candidates?" For the Cy Young Award, the answer appears to be yes.

A Mathematical Model to Predict Award Winners

Rebecca L. Sparks and David L. Abrahamson
Rhode Island College

Baseball has always been a delight for the mathematically inclined. Numerical records have been integral to following the sport, and popular measurements like batting averages have enticed many youngsters to experiment with simple calculations. In the last three decades, a great many fans and even some baseball executives have brought significant mathematical sophistication and the scientific method to the study of baseball. Evidence of this movement appears in publications ranging from the daily sports page to this magazine. We know that a player's offensive contribution can be measured by things such as his slugging and on-base averages; we know that a team's won-lost record can be predicted from its totals of runs scored and runs allowed. There have also been many discussions on using mathematics to compare players across eras. Organizations like the Society for American Baseball Research (SABR) have welcomed mathematical analyses; indeed, the name SABR has given rise to the term "sabermetrics" to describe the mathematical study of the various facets of the game.

In this article, we wish to follow a somewhat uncommon direction. Rather than analyzing how some measures of on-field performance correlate with other occurrences within the game, we want to see how they predict an off-field assessment. When a ballot is cast for an award—the Rookie of the Year, the Most Valuable Player, and other such awards—the voter presumably has some criteria in mind. One voter may pay particular attention to batting average, one may emphasize runs batted in, one may look primarily at where a player's team finished in the standings, and so on. A voter may or may not publish his criteria; indeed, he may very well be unaware of exactly what those factors are as he forms his impressions of various players. When the ballots are tallied, the voters will have ranked the candidates for an award, from first place on down. Every year there are an untold number of spirited discussions of how voting will go, with predictions from all corners about who will win.

One cannot help but wonder if a voting result is in fact predictable from the data available to the voters. Is it possible to look at the information for each candidate and combine it in such a way that we can predict how the voters will compare him to the others? In at least one case, we believe the answer is in the affirmative. In what follows, we will present a way to predict the ranking of candidates for one major award.

The Cy Young Award

In the 1950s, baseball created the Cy Young Award to honor the most outstanding pitcher in major league baseball. (The award is named after the legendary pitcher Denton True "Cy" Young, from the early era of baseball. His nickname came from batters who felt his fastball blow by them like a cyclone.) In the 1960s, the practice changed to provide each of the American and National leagues with an honoree, and it remains that way today. The award is voted upon by members of the Baseball Writers Association of America, two from each city in the league. Each writer in the electorate casts a preference ballot, specifying his choices for first, second, and third place, and the votes are tallied using a 5–3–1 point count (that is, 5 points for each first place vote, 3 points for each second place vote, and 1 point for each third place vote). The resulting point totals determine the final ranking of top pitchers within a league. The identities of the voters change frequently, but we will see that the voting results follow a predictable course.

Voters may use any criteria they wish. Perhaps some feel that a pitcher's number of wins is paramount, some others look first at earned run average, and others use still different statistics. There is room for a great deal of honest disagreement here. Let us look at the case of hypothetical stars John and Kevin. John plays for a pennant winner and has a won-lost

record of 24-8 with an earned run average of 2.95 while Kevin plays for a struggling team and has a record of 17-11 with an earned run average (ERA) of 1.89. Ignoring other statistics for a moment, who seems to be the better pitcher? Some might say that John is the man, as his 24 wins are a tremendous accomplishment. Others would vote for Kevin, citing his truly remarkable 1.89 ERA. Those in the latter group might say that Kevin's poorer won-lost record is not his fault—even if he pitches well, his team may score very few runs on offense—while those in the former group could say that wins are the bottom line and that if John was lucky enough to get some runs to work with, he turned them into wins and may not have even "needed" a microscopic ERA. An argument is at hand, even though this example cites only three statistics. (By the way, our hypothetical men are in fact John Smoltz and Kevin Brown of the 1996 National League. Smoltz blitzed Brown and the rest of the league in the award voting, receiving 136 points out of a possible 140.)

Our goal in this is not to argue how one *should* evaluate pitchers; although we have our own strong opinions on what statistics are most meaningful in describing a pitcher's effectiveness, we don't have a vote for the award. The task at hand is to try to predict how the actual award voting will go: can we take the season statistics for each pitcher in the league and come up with a formula that will predict exactly how the voters will rank those pitchers?

Before we get to the mathematics, we need to comment on an issue regarding the roles pitchers play. Over the long history of baseball, starting pitchers have been considered much more important than relief pitchers. Starters pitch many more innings over the course of a season and, for a century, teams made all their best pitchers starters. Bullpen dwellers were second-tier citizens on a pitching staff. But in the second half of the twentieth century, relievers have come to be considered much more important, and this emphasis continues to grow today. Fewer and fewer starting pitchers complete nine-inning games, and most teams have a few specialists in the bullpen, including one whose job is to slam the door on the opponent when his team has a small lead in the last inning. This "closer" is often an excellent pitcher and is something of a star to fans and team officials alike. Most people rate his effectiveness by how many saves he accumulates, and those numbers have mounted over the years as closers are used more often. (There is, however, significant disagreement about the actual relevance of the save as a meaningful statistic.) Relief pitchers, particularly closers, are taken more and more seriously by Cy Young Award voters as the years go by, but they still do not often place in the top spots in the voting. As a result, there is very little data regarding the appearance of relief pitchers near the top of voting results, and so we will restrict our model to starting pitchers. (As luck would have it, in 2003 the National League Cy Young Award winner was bullpen pitcher Eric Gagne. We refer to this situation as a "Cy of Relief".)

A Mathematical Model

For starting pitchers, the most commonly cited statistics are wins, losses, earned run average, and strikeouts, so we will include them in our model. With any award voting in baseball, extra attention seems to be paid to players on the league's better teams—there is a perception that if a player's team is a winner, he must be especially valuable and succeeding in pressure-packed situations—so we also include the winning percentage of each pitcher's team. We want a simple model, one whose parameters are familiar to all fans, so we'll use just those five statistics: wins (W), losses (L), earned run average (ERA), team winning percentage (TWP), and strikeouts (K).

For pitcher l in year j, we wish to calculate a score S_{lj} based on those five statistics. To make the final results easier to understand, we will put all those five on the same scale. We choose to put each on a scale of zero to ten by simple linear formulas, giving values p_{lj1} through p_{lj5} defined as follows:

$$\begin{aligned} p_{lj1} &= \frac{W}{3} \\ p_{lj2} &= 10\left(\frac{15-L}{15}\right) \\ p_{lj3} &= 12.5 - 2.5(ERA) \\ p_{lj4} &= 20(TWP - 0.25) \\ p_{lj5} &= 10\left(\frac{K-50}{333}\right). \end{aligned}$$

These formulas are chosen with the idea in mind that a parameter value of zero should correspond to a performance of no value to voters, while a value of ten should be historic in the modern era of baseball. If pitcher l in year j wins no games, then he has $p_{lj1} = 0$, while one who wins 30 games gets ten points. A starter with just 50 strikeouts earns a zero in the parameter p_{lj5}, while one who fans 383 to tie Nolan Ryan's record gets ten points, and so on. While negative values of some parameters are an algebraic possibility, it is not conceivable that a Cy Young Award contender will attain this dubious distinction. As an example, John Smoltz's 1996 season, mentioned above, had his 24 wins, 8 losses, ERA of 2.95, team winning proportion of 0.593, and 276 strikeouts convert to $(p_{lj1}, p_{lj2}, p_{lj3}, p_{lj4}, p_{lj5})$ = (8.0000, 4.6667, 5.1250, 6.8600, 6.7868).

We will then calculate each score S_{lj} as a weighted sum of these parameters:

$$S_{lj} = \sum_{k=1}^{5} x_k p_{ljk},$$

where the weights x_k are to be found in such a way that, for

every year j the scores S_{lj} mimic the order of finish in the award voting. We'll assign $l = 1$ to the winner of the award, $l = 2$ to the second-place finisher, and $l = 3$ to the pitcher finishing third in the vote, and the index j will vary from 1 to m if we are using m seasons worth of data in our work.

Historically, there are usually no more than three candidates who get serious consideration in the voting, so we will try to match just that portion of the voting results: the pitcher who wins the vote should have the highest score (S_{1j}) in the league that year, the second-place finisher should have the next best score (S_{2j}), and the third-place finisher should have the next best score (S_{3j}). The weights x_k are to be nonnegative, of course, and we further require that their sum be one. Requiring $\sum x_k = 1$ makes the score S_{lj} a convex combination of the parameters p_{lj1} through p_{lj5}, making it easy to compare the relative importance of their values by seeing how large a weight is assigned to each.

At this point, we have a formulation of the problem at hand. We wish to find numbers x_1 through x_5 such that all of the following are true:

$$\sum_{k=1}^{5} x_k = 1 \tag{1}$$

$$x_k \geq 0, \quad k = 1,\ldots,5 \tag{2}$$

$$S_{1j} > S_{2j} > S_{3j}, \quad j = 1,\ldots,m \tag{3}$$

It is worth writing out the meaning of the set of conditions in (3). They say that for each j we need

$$\sum_{k=1}^{5} x_k p_{1jk} > \sum_{k=1}^{5} x_k p_{2jk} > \sum_{k=1}^{5} x_k p_{3jk},$$

or

$$\sum_{k=1}^{5} x_k (p_{1jk} - p_{2jk}) > 0 \text{ and } \sum_{k=1}^{5} x_k (p_{2jk} - p_{3jk}) > 0 \; j = 1, \ldots, m. \tag{4}$$

It is possible to eliminate (1) by solving for one of the weights in terms of the others, but we will not need to do so. The other conditions, namely (2) and (4), are very nearly the sort of constraints which appear in a standard linear programming problem. If we replace (4) by a set of conditions in which the inequality is not strict, say by replacing zero by a small positive number, then (2) and (4) together will specify exactly the sort of set that appears as the feasible region in a linear programming problem.

Since we will set up our problem as a linear programming problem, it is necessary to choose an objective function that makes sense for this problem. Since we would like our model to predict the winner as it happens in real life, a natural one would be to maximize the scores of first-place finishers. That is, we would like to maximize S_{1j} for all years j in the data set. In order to take into account all the winners at the same time, we chose our objective function to be

$$F(x_1,\ldots,x_5) = \sum_{j=1}^{m} S_{1j} = \sum_{j=1}^{m}\left(\sum_{k=1}^{5} x_k p_{1jk}\right).$$

No doubt the reader could come up with other objective functions, but we suggest this one since it seems most reasonable from a baseball point of view.

Thus, the problem we will attempt to solve is as follows:

Problem (LP-Cy).

Given $a > 0$, find $x = (x_1, \ldots, x_5)$ that solves the following linear programming problem:

$$\text{Maximize:} \quad F(x) = \sum_{j=1}^{m} S_{1j} = \sum_{j=1}^{m}\left(\sum_{k=1}^{5} x_k p_{1jk}\right).$$

$$\begin{aligned}
\text{Subject to:} \quad & \sum_{k=1}^{5} x_k (p_{1jk} - p_{2jk}) \geq a, \; j = 1,\ldots,m \\
& \sum_{k=1}^{5} x_k (p_{2jk} - p_{3jk}) \geq a, \; j = 1,\ldots,m \\
& \sum_{k=1}^{5} x_k = 1 \\
& x_k \geq 0, \quad k = 1,\ldots,5.
\end{aligned}$$

Finding a solution to (LP-Cy)

The data used were from both the American and National leagues for the seasons 1993 through 2002 ($m = 20$). In an attempt to separate scores (thereby avoiding ties for first place), we set $a = 0.001$, and used the linear programming package available within *Mathematica* to find solutions to the problem.

Using the data described above, no feasible solution exists. Is all our work for naught? Not at all. After a closer look at the 20 seasons' worth of data, a possible reason appears. In 1995 in the American League, the second-place finisher in the award voting was David Cone of the New York Yankees, with an 18-8 record, 3.57 ERA, 191 strikeouts, and a team win proportion of 0.549, while Mike Mussina of the Baltimore Orioles finished third with 19 wins, 9 losses, a 3.29 ERA, 158 strikeouts, and a team win proportion of 0.493. Although Cone's statistics look a little better than Mussina's in some categories, his fewer wins and worse ERA led us to wonder if the snag in the problem might be that Cone is supposed to have a lower score than Mussina in 1995. We chose to delete that constraint, and then tried again. With that lone constraint deleted, the problem has a solution. The resulting values of the weights are as follows:

$$\begin{aligned}
x_1 &= 0.578084, \text{ (Wins)} \\
x_2 &= 0.00999357, \text{ (Losses)} \\
x_3 &= 0.196700, \text{ (ERA)} \\
x_4 &= 0.0784757, \text{ (Team Winning Proportion)} \\
x_5 &= 0.136747, \text{ (Strikeouts)}
\end{aligned}$$

These weights tell us that the voters take the pitchers' win totals much more seriously than any other category, as the weight of 0.578084 on that category indicates. Earned run average is the next most important to the voters, followed by strikeouts. Team winning percentage plays a smaller but nontrivial role, and losses seem nearly irrelevant to the voters.

Now, if we wish to find the top three finishers (among starting pitchers) in the Cy Young balloting in any of the seasons considered, we just have to take each pitcher's vector of values on the ten-point scales and find its dot product with the solution vector found above. The pitcher with the highest result wins the award, the pitcher with the second-highest result finishes second, and the pitcher with the third-highest result finishes third. Going back to our John Smoltz example from 1996, his score for that year is: $(0.578084, 0.00999357, 0.196700, 0.0784757, 0.136747)\cdot(8.0000, 4.6667, 5.1250, 6.8600, 6.7868) = 7.14581$.

The bottom line is that the score for each pitcher is a weighted average of his (ten-point converted) wins, losses, ERA, team winning proportion, and strikeouts.

From scores calculated using this new method, our "predicted" top three finishers precisely match the actual top three finishers every year in both leagues with the exception of the 1995 American League. In this case it is important to note, however, that our model correctly predicts the winner of the Cy Young award that year; it only fails to agree on who should finish in second and third place.

To help the reader get a feel for the method and the resulting scores, we summarize our results in Table 1.

	Winner	Second place	Third place
A.L. 1993	Jack McDowell 6.03559	Randy Johnson 6.03459	Kevin Appier 5.6957
A.L. 1994*	David Cone 4.98132	Jimmy Key 4.97198	Randy Johnson 4.38253
A.L. 1995*	Randy Johnson 6.25932	David Cone 5.26665	Mike Mussina 5.36696
A.L. 1996	Pat Hentgen 5.60894	Andy Pettitte 5.60794	Charles Nagy 5.17765
A.L. 1997	Roger Clemens 6.88805	Randy Johnson 6.73468	Brad Radke 5.21891
A.L. 1998	Roger Clemens 6.43687	Pedro Martinez 6.07660	David Wells 5.46093
A.L. 1999	Pedro Martinez 7.54405	Mike Mussina 5.12298	Bartolo Colon 5.05505
A.L. 2000	Pedro Martinez 6.52411	Tim Hudson 5.31983	David Wells 5.22084
A.L. 2001	Roger Clemens 5.87582	Mark Mulder 5.87482	Freddy Garcia 5.68281
A.L. 2002	Barry Zito 6.75293	Pedro Martinez 6.55923	Derek Lowe 6.10798
N.L. 1993	Greg Maddux 6.40590	Bill Swift 6.21047	Tom Glavine 6.08867
N.L. 1994*	Greg Maddux 5.81304	Ken Hill 4.74584	Bret Saberhagen 4.63625
N.L. 1995*	Greg Maddux 6.53153	Pete Schourek 5.38247	Tom Glavine 4.98534
N.L. 1996	John Smoltz 7.14581	Kevin Brown 5.66237	Andy Benes 4.98875
N.L. 1997	Pedro Martinez 6.25659	Greg Maddux 6.21342	Denny Neagle 6.00519
N.L. 1998	Tom Glavine 6.23146	Kevin Brown 6.21741	Greg Maddux 6.14203
N.L. 1999	Randy Johnson 6.41553	Mike Hampton 6.41453	Kevin Millwood 5.90500
N.L. 2000	Randy Johnson 6.52628	Tom Glavine 5.81958	Greg Maddux 5.78694
N.L. 2001	Randy Johnson 7.15283	Curt Schilling 6.78015	Matt Morris 6.25364
N.L. 2002	Randy Johnson 7.73261	Curt Schilling 7.00518	Roy Oswalt 5.75078

Table 1. Top Cy Young Award Vote-Getters and Their Associated Scores

*A labor dispute shortened the 1994 and 1995 seasons to significantly fewer than the usual 162 games. This lowers overall scores somewhat in those seasons by decreasing the opportunities for wins and strikeouts.

We can interpret the scores in Table 1 further. Our suspicions about the 1995 American League voting were correct. Considering all the data except the Cone/Mussina constraint, the voters perform consistently and weigh wins the most strongly and ERA second. That season, though, Cone finishes ahead of Mussina despite trailing in both of these key categories. The voting pattern that year has an inconsistency in it. (One might speculate that the quirk has to do with the vast publicity garnered by New York teams.)

Since a "10" in any category indicates a historic achievement, it would be most unlikely to have any pitcher attain a score near that value. The two highest scores which occur during the seasons of our study are Randy Johnson's 7.73261 in the 2002 N.L., and Pedro Martinez's 7.54405 in the 1999 A.L.; readers may recall the terrifying dominance of those two pitchers.

As noted above, the model presented here used data from 1993 to 2002. With two more baseball seasons now complete, one might ask how the model's formula fared in predicting further outcomes. Applying the formula derived above to the 2003 and 2004 American League seasons correctly predicts the first-place and second-place finishers in the voting for both seasons. (In 2004, only two starting pitchers received any points in the voting, and in 2003, the model's predictions invert the third-place and fourth-place finishers in the voting.) The 2003 National League award went to closer Eric Gagne for his astonishing relief work, so the model does not apply to that season. The 2004 National League season presented a snag for the model, as it predicted that Houston's 20-game winner Roy Oswalt would edge teammate Roger Clemens in the voting. The award actually went to the renowned Clemens and his 18 wins. It seems possible that the intense media coverage Clemens received in coming back from a brief "retirement" may have played a role with the voters.

Beyond Baseball

Can our constrained-optimization approach be used on similar problems? For the baseball fan, other awards are certainly a source of questions, but the issue of comparing pitchers with

Continued on page 13

In practice, jumpers can control the probability of fouling by choosing an "aiming line" that is some distance (on the order of a few centimeters) before the official takeoff line. The closer the aiming line is to the takeoff line, the greater the chance of fouling. However, the length of the jump is measured from the takeoff line, so choosing an aiming line that is too far from the takeoff line decreases the net length of the jump.

Other Ways of Scoring

Using a best-of-three-attempts scoring system is not the only possibility for events of this type. One alternative would be to add (or average) the three attempts. This would allow a jumper to compensate for a poor jump with an exceptionally long jump, perhaps by aiming closer to the official takeoff line. Such a scoring scheme is used in many golf tournaments, where a golfer's score for the tournament is the sum of the scores in the four rounds. (The lowest total wins.) It is fairly common that the golfer with the best round of the tournament does not win. Consistency is more important.

Let's see what would happen if such a system was used in the long jump. Let $S = X_1 + X_2 + X_3$ be the sum of the three jumps. If the jumper does not foul, then $E[S] = 3\mu_X$. However, if the jumper fouls once, then $E[S] = 2\mu_X$, a substantial reduction. In contrast, under the current system, a single foul reduces the expected score from $.846\sigma_X + \mu_X$ to $.564\sigma_X + \mu_X$, a much less significant effect (depending on the jumper's consistency).

If θ is the probability of a successful jump, then conditioning on the number of successful jumps, we have $E[S] = \theta^3 \cdot (3\mu_X) + 3\,\theta^2(1-\theta)(2\mu_X) + 3\theta(1-\theta)^2\mu_X = 3\mu_X\theta$. So Allison's expected score reduces by 10% as θ decreases from 1 to .9. Under the current system, the decrease is 0.5%. (Note: We could also derive this by observing that $E[X_i] = \mu_X\theta$ for each i. Then $E[S] = E[X_1] + E[X_2] + E[X_3] = 3\mu_X\theta$.)

In essence, using the sum of the three jumps would eliminate any jumper who fouls from having any chance of winning against a jumper who doesn't foul.

We have, of course, assumed that a foul counts as a jump of length 0. Alternatively, we could say that a foul results in some distance (perhaps 25 cm) that is deducted from the length of the jump. This would be similar to what happens in some competitions, such as obstacle courses, where failure to complete some aspect of the course results in a "time penalty" added to the final time. In golf, certain offenses result in a one- or two-stroke penalty, rather than having to replay the entire hole, or forfeiting the round. Now if we add the lengths of the jumps with this penalty scheme, a jumper could possibly compensate for a foul by executing two outstanding jumps on the other attempts. ■

Further Reading

Shaul P. Ladany and Georghios P. Sphicas, Dynamic Policies in the Long Jump, in *Management Science in Sports*, S. P. Ladany, R.E. Machol, and D. G. Morrison, eds., North-Holland Publishing, 1976.

Continued from page 8

position players has been a source of disagreements among voters and might therefore give rise to some unpredictable voting. Studies of awards in basketball sound promising, for all players accumulate statistics in essentially the same categories. The same might be said for hockey if goalies are not under consideration.

If not sports, then what? One example may arise in business or even academia. Companies are usually evaluated on a regular basis, and they may be ranked regarding earnings, size, speed of shipping products, affordability, and quality of their products. "Top Colleges" are ranked by comparing similar categories. The moral is always the same for the mathematical modeler: more often than we may know, there is a pattern out there. We just have to keep thinking creatively, and we have got a good chance of finding it. ■

Further Reading

D. L. Abrahamson and H. Salzberg, A Mathematical Modeling Perspective on Baseball Statistics, *Mathematics in College*, 1998.

B. James, *The Bill James Historical Baseball Abstract*, New York: Villard Books, 1985.

J. Thorn and P. Palmer, *The Hidden Game of Baseball*, Garden City, New York: 1985.