2014

Predicting the Cy Young Award Winner

Stephen Ockerman  
*Western Oregon University*, sockerman10@mail.wou.edu

Matthew Nabity  
*Western Oregon University*, nabitym@wou.edu

Follow this and additional works at: [https://digitalcommons.wou.edu/pure](https://digitalcommons.wou.edu/pure)  
Part of the [Numerical Analysis and Computation Commons](https://digitalcommons.wou.edu/pure)

**Recommended Citation**  
Ockerman, Stephen and Nabity, Matthew (2014) "Predicting the Cy Young Award Winner," *PURE Insights*: Vol. 3, Article 9.  
Available at: [https://digitalcommons.wou.edu/pure/vol3/iss1/9](https://digitalcommons.wou.edu/pure/vol3/iss1/9)

This Article is brought to you for free and open access by the Student Scholarship at Digital Commons@WOU. It has been accepted for inclusion in PURE Insights by an authorized editor of Digital Commons@WOU. For more information, please contact digitalcommons@wou.edu.
Predicting the Cy Young Award Winner

Abstract
Here we examine the application of a decision model to predicting the winner of the Cy Young Award. We investigate a current model cast as a linear programming problem and explore its ability to correctly predict award winners for recent seasons of professional baseball. We suggest the addition of another baseball statistic which leads to a new model. We explore the success of both models with numerical experiments and discuss the results.

Keywords
linear programming, mathematical model, sabermetrics
Predicting the Cy Young Award Winner

Stephen Ockerman and Matthew Nabity

1. Introduction

The game of baseball has long been associated with collecting and analyzing empirical data. Since the creation of Major League Baseball in 1903, various numerical measurements have been recorded and invented.

These records have been used by fans of the game to assess teams and players, and more recently, by those interested in further studying the game. The mathematical study of baseball statistics is often referred to as sabermetrics, named after the Society for American Baseball Research (SABR) [3]. The practice of sabermetrics is depicted in the book Moneyball and in the recent movie adaptation. Mathematical analysis of the many facets of the game of baseball has been growing steadily in recent years.

In 2005, a mathematical model to predict the winner of MLB’s Cy Young Award was suggested by Sparks and Abrahamson [5]. This model used a different approach than the methodologies common to sabermetrics. The authors attempted to use on-field statistics to forecast off-field assessments. The Cy Young Award is awarded to the most outstanding pitcher in each of the American and National Leagues, and the winners are chosen by voting members of the Baseball Writers Association of America. The mathematical model was formulated not to predict who should be awarded the Cy Young Award but to accurately predict how the award voting would go. That is, the model hoped to use the current season of statistics to predict how voters would rank the pitchers.

Data from the 1993 season through the 2002 season was used, specifically five common statistics: wins, losses, earned run average, team winning percentage, and strikeouts. Weights for a weighted average were determined by formulating and solving a linear programming problem. A linear programming problem has a linear function of the unknowns. The objective is to maximize or minimize this function subject to constraints that are also linear. In this model, weights for each of the statistics were determined and then used to compute a numerical score for each player. The player with the highest score was expected to win the award, the player with the second highest score would finish in second place, and the player with the next highest score would finish in third place.

The model using all 20 seasons of data did not have a solution. After a closer inspection of the data, the authors removed the statistics from the American League (AL) in 1995. With this single constraint removed, the model correctly predicted the voters’ choice for the top three finishers in each league in every year except for the AL in 1995. In this isolated case, the winner was correctly identified, but second and third places were not.

Nearly a decade later, this work began with a central question, does this model accurately capture the attitude of the voters today? Those that follow the game of baseball may be aware of the numerous statistics available and the often intense debates about which ones matter more for in-season performance and post season accolades. Many baseball fans may also be aware of recent emphasis on statistics such as WHIP, walks and hits per inning pitched, or WAR, wins above replacement.

To explore the relevance of the model, we revisit the formulation of Sparks and Abrahamson’s model and apply it to more recent seasons, namely the 2005 through 2013 seasons. Based on
the numerical results, we explore the addition of another statistic and suggest an updated version of the model. We report numerical results of our new model and discuss the results of our modeling efforts.

2. The Mathematical Model

When examining an award for pitchers, we need to understand the position. There are two major types of pitchers. The first, the starting pitcher, typically begins the game and pitches until relieved. The second, the relief pitcher, is any player that is not in the starting rotation. In recent years, the work of relief pitchers, specifically those that finish the game, has been increasingly appreciated by Cy Young voters. In fact, the National League (NL) Cy Young winner in 2003, Éric Gagné, was such a relief pitcher, often called a closer. Closers are usually judged by different standards than starting pitchers.

The model developed by Sparks and Abrahamson does not apply to the 2003 season in the NL as they restricted their analysis to include only starting pitchers. For the 2005 season, the model correctly predicted Chris Carpenter for the NL winner. In the AL that year, the consensus was that there was no stand out performance and many believed Mariano Rivera, a relief pitcher, would win. The mathematical model correctly predicted Bartolo Colón would be the AL winner, but Mariano Rivera, who finished second in the voting, was not included in the analysis as he was not a starting pitcher.

Other attempts to predict the Cy Young Award winner have been made using different mathematical techniques, for example the data mining approach by Smith et al [4]. Using a Bayesian classifier, the authors examined data from the years 1967 to 2006 and were more than 80% correct when restricting their analysis to only starting pitchers [4]. Accuracy suffered when including relief pitchers. Due to the difficult nature of including relief pitchers, the most successful models currently consider only starting pitchers.

The statistics used in the original model include wins (W), losses (L), earned run average (ERA), strikeouts (K) and team winning percentage (TWP). The first four of these measurements are fairly common, but TWP is not necessarily a direct assessment of an individual. One of the modeling assumptions is that players on better teams get more exposure and potentially more credit for the success of the team. To make the data easier to compare, Sparks and Abrahamson put all five statistics on the same scale, zero to ten, using simple linear transformations. The parameters were chosen so that a score near ten reflects a historic performance and a score near zero reflects a performance of little interest to voters. For pitcher \( i \) in year \( j \), they defined the following:

\[
p_{ij1} = \frac{W}{3} \tag{2.1}
\]

\[
p_{ij2} = 10 \left( \frac{15 - L}{15} \right) \tag{2.2}
\]

\[
p_{ij3} = 12.5 - 2.5(ERA) \tag{2.3}
\]

\[
p_{ij4} = 20(TWP - 0.25) \tag{2.4}
\]

\[
p_{ij5} = 10 \left( \frac{K - 50}{333} \right) \tag{2.5}
\]

Using the scaled data, a score for pitcher \( i \) in year \( j \), \( S_{ij} \), can be compute as weighted sum of these parameters:

\[
S_{ij} = \sum_{k=1}^{5} x_k p_{ijk},
\]

where the \( x_k \) are to be determined so that the pitcher that wins has the highest score in the league for year \( j \), the second-place finisher should have the second highest score, and the third-place finisher should have the third highest score.

Sparks and Abrahamson required that the nonnegative weights add up to one so that each score \( S_{ij} \) was a convex combination of the parameters \( p_{ijk}, k = 1, 2, \ldots, 5 \). The formulation thus far is to find numbers \( x_1 \) through \( x_5 \) so that all of the following are true:

\[
\sum_{k=1}^{5} x_k = 1 \tag{2.6}
\]

\[
x_k \geq 0, \; k = 1, \ldots, 5 \tag{2.7}
\]

\[
S_{1j} > S_{2j} > S_{3j}, \; j = 1, \ldots, m, \tag{2.8}
\]
where \( m \) is the number of seasons used. The constraints 2.7 and 2.8 are close to the types of constraints that appear in a linear programming problem. A linear programming problem is characterized by linear functions of the unknowns and linear inequalities and equalities of the constraints [2]. The goal is to maximize or minimize a specific objective subject to certain constraints. For example, if \( p_1 \) and \( p_2 \) are two measures of performance described above, and the goal is to find weights \( w_1 \) and \( w_2 \) that would maximize the weighted average \( w_1 p_1 + w_2 p_2 \), then the overall problem could be expressed as

\[
\text{Maximize:} \quad w_1 p_1 + w_2 p_2 \\
\text{Subject to:} \quad p_1 + p_2 \leq b, \quad p_1 \geq 0, p_2 \geq 0,
\]

where \( b \) is some number derived from the context. A successful solution to this simple linear program would compute values for \( w_1 \) and \( w_2 \). For details on linear programming problems and related algorithms, see [2]. Examining 2.8 more closely, we see that for each year \( j \)

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

or rearranging terms we have constraints of the form

\[
\sum_{k=1}^{5} x_k (p_{1jk} - p_{2jk}) > 0, j = 1, 2, \ldots, m, \\
\sum_{k=1}^{5} x_k (p_{2jk} - p_{3jk}) > 0, j = 1, 2, \ldots, m. \quad (2.9)
\]

The authors made these inequalities not strict by replacing zero with a small positive number. This adjustment made it so that constraints 2.7 and 2.9 specify the feasible region for a linear programming problem. Mathematically speaking, all that was needed now was something to optimize, that is, an objective function.

Sparks and Abrahamson set up a linear programming problem in which the score \( S_{1j} \) for all years \( j \) in the data set was maximized. To accomplish this, they chose to maximize the sum of all winners over the years in the data set. In linear programming terms, this was selected as the objective function for the maximization problem. The final form of the problem was as follows:

**Problem (LP1-CY)**

Given \( a > 0 \), find \( x = (x_1, \ldots, x_5) \) that satisfies:

\[
\text{Maximize:} \quad F(x) = \sum_{j=1}^{m} S_{1j} = \sum_{j=1}^{m} \sum_{k=1}^{5} x_k p_{1jk} \\
\text{subject to:} \quad \sum_{k=1}^{5} x_k (p_{1jk} - p_{2jk}) \geq a, \quad j = 1, \ldots, m \quad (2.11) \\
\sum_{k=1}^{5} x_k (p_{2jk} - p_{3jk}) \geq a, \quad j = 1, \ldots, m \quad (2.12) \\
\sum_{k=1}^{5} x_k = 1 \quad (2.13) \\
x_k \geq 0, \quad k = 1, \ldots, 5. \quad (2.14)
\]

We note that when using data from all 20 seasons, no feasible solution was found to exist. The 1995 season caused problems for the model LP1-CY and the authors chose to delete the AL information from that year. Removing this single constraint allowed the linear programming package from Mathematica to compute the following weights:

\[
x_1 = 0.578084, \quad (W) \\
x_2 = 0.00999357, \quad (L) \\
x_3 = 0.197600, \quad (ERA) \\
x_4 = 0.0784757, \quad (TWP) \\
x_5 = 0.136747, \quad (K).
\]

These weights indicated that for the seasons under consideration, total wins was the most important category, followed by ERA and then by strikeouts. The assumption that TWP played a role was somewhat validated by the result that it was more important than total losses, which was practically irrelevant relative to the other components.
2.1. Numerical Results Part One

To explore the relevance of this model on more recent seasons, we performed a few numerical experiments. All computations were done using the linear programming capabilities of standard functions in MATLAB R 2013a. First, we used the original model formulation but only constraints from the past nine seasons, 2005 to 2013. As in the initial attempt by the original authors, no feasible solutions were found. Recall that the original authors had to remove a constraint, namely the 1995 season data from the AL. This may have been easy to identify as there was a players strike that ended the 1994 season and carried on into the 1995 season. When looking at more recent seasons, we had no obvious seasons to look at.

To gain some insight into the most recent nine seasons of data, we computed the scores for the top three finishers using the original weights computed by Sparks and Abrahamson for the data from 1993 through 2002. We found that the overall winner in the AL was correctly identified in six of the nine years: 2005, 2006, 2008, 2011, 2012, and 2013. Of these, the model did not properly account for relief pitcher Mariano Rivera in 2005, and had the second place and third place finishers in the wrong order in 2012. The story was about the same for the results in the NL. The model correctly identified the top three finishers in five of the nine years: 2005, 2006, 2007, 2010, and 2011.

Collectively, the weights determined by the LP1-CY were only successful in predicting the winners in both leagues in 2006 and 2011. We were unable to identify a pattern for the success or failure of the model, and there were no obvious seasons to consider removing from the set of constraints. Based on the assumption that the voters’ attitudes have been changing in recent years, we set out to incorporate additional information.

3. A New Model

Though there have been successful relief pitchers lately, we also opt to restrict our analysis to starting pitchers. Despite the fact that the weights computed by the original authors suggest that the number of losses seems unimportant to voters, we base our model on the five original statistics and choose to include an additional statistic. Walks plus hits per inning pitched (WHIP) is a sabermetric measurement that has been used to assess pitchers for over three decades. In recent years this statistic has found its way into MLB box scores on popular sports websites. The measurement attempts to measure a pitcher’s effectiveness against batters. The lowest single-season WHIP in MLB history, 0.7373, was recorded by Pedro Martinez during the 2000 season while playing for the Boston Red Sox [1]. Using this value as a historic performance, we define the transformation

\[ p_{ij6} = 10 (2 - WHIP) - 2.627, \]

(3.1)

to incorporate WHIP into the model based on LP1-CY. Here a WHIP of 0.7373 would score ten points. Adding this component to the data and using the scaled data from LP1-CY, we now consider the weighted sum or objective function

\[ S_{ij} = \sum_{k=1}^{6} x_k p_{ijk}, \]

where the \( x_k \) are to be determined so that the pitcher that wins again has the highest score. Reformulating this as a linear programming problem in the same manner as before, we have the follow-
Problem (LP2-CY)

Given \( a > 0 \), find \( x = (x_1, \ldots, x_6) \) that satisfies:

Maximize: \( F(x) = \sum_{j=1}^{m} S_{1j} \) \hspace{1cm} (3.2)

\[
= \sum_{j=1}^{m} \sum_{k=1}^{6} x_k p_{1jk}
\]

subject to:

\[
\sum_{k=1}^{6} x_k (p_{1jk} - p_{2jk}) \geq a, \quad j = 1, \ldots, m \quad (3.3)
\]

\[
\sum_{k=1}^{6} x_k (p_{2jk} - p_{3jk}) \geq a, \quad j = 1, \ldots, m \quad (3.4)
\]

\[
\sum_{k=1}^{6} x_k = 1 \quad (3.5)
\]

\( x_k \geq 0, \quad k = 1, \ldots, 6. \) \hspace{1cm} (3.6)

The incorporation of an additional measurement changes both the objective function and the constraints that define the feasible region. We now seek six weights to help capture the voters’ attitude.

3.1. Numerical Results Part Two

In this section we report the results of further numerical experiments using both the original model LP1-CY and our updated version LP2-CY. Here we examine solutions to each of the models for various sets of constraints. The goal is to identify weights that most accurately predict the top three finishers.

We began with data from both leagues for the most recent seasons, 2005 through 2013. Recall from the previous numerical experiments, there was no feasible solution to LP1-CY using data from these nine seasons. We observed the same for our new model LP2-CY using these same constraints. To investigate this further, we turned to the original weights computed using the 1993 through 2002 seasons, excluding the AL results from 1995. Looking at the NL results using these weights, we noticed that there seemed to be a change after the 2007 season. This motivated us to restrict our constraints to data from the most recent six seasons.

**Experiment 1**

Here we used data from the past 6 seasons, 2008 through 2013, for both leagues. For LP1-CY, we computed the weights to be

\[
x_1 = 0.000000, \quad (W)
\]

\[
x_2 = 0.051184, \quad (L)
\]

\[
x_3 = 0.780348, \quad (ERA)
\]

\[
x_4 = 0.083859, \quad (TWP)
\]

\[
x_5 = 0.084608, \quad (K),
\]

and for LP2-CY we found the weights to be

\[
x_1 = 0.000000, \quad (W)
\]

\[
x_2 = 0.051184, \quad (L)
\]

\[
x_3 = 0.780348, \quad (ERA)
\]

\[
x_4 = 0.083859, \quad (TWP)
\]

\[
x_5 = 0.084608, \quad (K),
\]

\[
x_6 = 0.000000, \quad (WHIP).
\]

We found the weights to be the same for either model as the weight for WHIP was determined to be \( x_6 = 0 \). Restricting our analysis to the seasons 2008 through 2013 seems to indicate that our new statistic may be extraneous. Here wins and WHIP do not seem to be factors, and ERA is the main component emphasized.

To assess the performance of these weights, we compute the numerical rankings for each of the top three finishers for both leagues. Table 3.1 shows the actual top three finishers in the AL and the scores computed by both models with incorrect predictions in red. We see that these weights were rather successful as all the top three finishers in the AL were correctly identified. Turning to the NL, we see quite a different story. Table 3.2 displays the actual top three finishers with the scores computed using weights identified by both models. Again, incorrect scores are highlighted in red. Here we see that the NL winner was only correctly
identified half the time and all top three finishers were correctly identified in only two of the six years. While we were able to compute a solution to both LP-CY1 and LP-CY2, the weights are not performing well and we opt to further explore the constraints.

**EXPERIMENT 2**

We now restrict the constraints to data from the 2009 through 2013 seasons for both leagues. That is, we removed the statistics from the 2008 season. Using only the most recent five seasons, LP1-CY produced the same weights as in the first experiment. The scores for the top three finishers are they same as those reported in Table 3.1 and

Table 3.2. We saw that these weights correctly identified all top three finishers in the AL but were much less successful in the NL, especially for the two most recent seasons as illustrated in Table 3.2.

Using LP2-CY with the constraints from the 2009 through 2013 seasons, we found the weights to be

\[
x_1 = 0.301385, \quad (W)
\]
\[
x_2 = 0.048033, \quad (L)
\]
\[
x_3 = 0.000000, \quad (ERA)
\]
\[
x_4 = 0.000000, \quad (TWP)
\]
\[
x_5 = 0.197455, \quad (K)
\]
\[
x_6 = 0.453127, \quad (WHIP).
\]

Now WHIP and wins are more important components, whereas both ERA and TWP are nonfactors. Using these weights, we compute the scores for the top three finishers in each league and compare the performance to the results in the first experiment. Table 3.3 shows the results for the AL using the weights computed by LP2-CY for the second experiment. Again, incorrect predictions are highlighted in red.

Table 3.3: AL Top Cy Young Finishers and Associated Scores, experiment 2

<table>
<thead>
<tr>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>'09</td>
<td>Z. Greinke</td>
<td>F. Hernandez</td>
</tr>
<tr>
<td>5.9801</td>
<td>5.9484</td>
<td>5.9474</td>
</tr>
<tr>
<td>'10</td>
<td>F. Hernandez</td>
<td>D. Price</td>
</tr>
<tr>
<td>5.5639</td>
<td>5.4816</td>
<td>5.7130</td>
</tr>
<tr>
<td>'11</td>
<td>J. Verlander</td>
<td>J. Weaver</td>
</tr>
<tr>
<td>7.6206</td>
<td>6.2056</td>
<td>5.8872</td>
</tr>
<tr>
<td>'12</td>
<td>D. Price</td>
<td>J. Verlander</td>
</tr>
<tr>
<td>6.1363</td>
<td>6.1353</td>
<td>6.1343</td>
</tr>
<tr>
<td>'13</td>
<td>M. Scherzer</td>
<td>Y. Darvish</td>
</tr>
<tr>
<td>7.0974</td>
<td>5.8543</td>
<td>5.8089</td>
</tr>
</tbody>
</table>
season. If these scores are compared to the scores calculated in Table 3.3, we see that Max Scherzer had a higher score in Table 3.3. This suggests that the addition of WHIP may have been helpful for seasons such as this when players have comparable statistics.

Unfortunately, the weights computed by LP2-CY failed to correctly identify any of the finishers for the 2010 season. We will investigate this further in ensuing experiments. To fully assess the performance of our model we turn our attention to the NL results. Table 3.4 shows the scores for the top three NL finishers when using weights computed by LP2-CY and only data from the 2009 through 2013 season. Here we see a much different story than for the AL results.

We can look at Table 3.2 for the performance of LP1-CY in the second experiment as the computed weights remained the same. These weights correctly predicted the winner in only three of the five years, 2010, 2011, and 2013, and correctly identified all three finishers only in 2010 and 2011. For comparison, the weights computed by LP2-CY accurately captured all first place finishers over the years in question. The only incorrect score occurred in 2012 where second and third place were out of order. This is a significant improvement from the results in Table 3.2 and may offset the issue with the 2010 season in the AL. When examining the results from both leagues, it seems that for the seasons under consideration, LP2-CY more accurately predicts not only the winner but also the top three finishers.

Additional Experiments

We performed several other experiments to attempt to identify problematic seasons. We found no feasible solutions to both LP1-CY and LP2-CY for the following sets of constraints: all data from 2005 through 2013, all data from 2005 through 2013 when omitting 2010 AL statistics, all data from 2005 through 2013 omitting all 2010 data, all data from 2005 to 2013 omitting 2010 AL statistics and 2012 NL statistics, all data from seasons 2005 through 2008, and all data from 2009 through 2013 omitting 2010 AL statistics.

We found two sets of constraints that generated two different sets of weights for LP2-CY when there was no feasible solution to LP1-CY. These configurations included the 2009 through 2013 seasons when omitting the 2010 AL data and the 2012 NL data and the 2009 through 2013 seasons when omitting the 2012 NL data. In either case, the weights switched the second place and third place finishers in the NL in 2012 as we saw in the second experiment in Table 3.4. The success of these other weights in the AL was not as good as what we observed in the second experiment in Table 3.3.

4. Discussion of Results

In this work, we examined the issue of predicting the Cy Young Award winner using regular season statistics and a decision model cast as a linear program. We applied an existing model, LP1-CY, to current seasons in an effort to see if the original statistics were enough to correctly predict the award winners. We found that when applied to recent seasons, the model forecast suffered. To investigate the addition of another statistic, we updated the model to include the sabermetric measurement WHIP. When restricted to data from seasons 2009 through 2013, the new model LP2-CY was much more successful in accurately predicting the top three finishers. Our model did have difficulty with the 2010 season in the AL and switched the second place and third place finishers in 2012. Having successfully made the case for the addition of
WHIP, we plan to further investigate refinements to LP2-CY. A closer examination of individual seasons in both leagues may shed some light on problematic constraints or seasons. Additionally, the incorporation of additional sabermetric measurements may help better capture the voters’ behavior. We plan to investigate possible model improvements in time for the coming postseason.

References


