I. Research Objective

Speculating about which pitcher will win the Cy Young Award has long been a pastime of baseball fans. In this paper, we identify which metrics affect a pitcher’s chances of winning the Cy Young Award, and the marginal effect of each metric. Our results were found using an ordinary least squares regression with a data set containing all pitchers in the American league who received at least one vote for the Cy Young award between 1970 - 2009. Our results show that voters favor pitchers with a high number of wins and a strong strikeout rate. Starting pitchers are also heavily favored over relievers.

IV. Data

Cross-sectional data set containing all pitchers in the American League that received a vote for the Cy Young Award between 1970 and 2009

Sample size: 298

Data Sources:
- Most player data came from Fangraphs.com
- Data on Cy Young votes came from BaseballReference.com

Data Challenges:
- Identifying pitchers as starters or relievers
  - Neither of the data sources indicated whether a pitcher was a starter or reliever
  - Pitchers were identified as starters or relievers based on their games played, games started, and saves

II. Empirical Model and Variables


Log(Wins) = Logarithm of the percentage of the total Cy Young votes possible in that given season.

Wns = Number of wins earned by the pitcher in that given season.

Losses = Number of losses attributed to the pitcher in that given season.

Saves = Number of saves earned by the pitcher in that given season.

IP = Number of innings pitched by the pitcher in that given season.

ERA = Number of runs given up per nine innings pitched by the pitcher in that given season.

K/IP = Number of strikeouts earned by the pitcher in that given season multiplied by the number of innings that they pitched.

BB/IP = Number of walks issued by the pitcher in that given season multiplied by the number of innings that they pitched.

HR/IP = Number of home runs hit up by the pitcher in that given season multiplied by the number of innings that they pitched.

Starter = Dummy variable for whether the pitcher was a starter or reliever.

* Denotes player whose name = 1 - 298

III. Hypotheses

Wns, is hypothesized to have a positive impact on Log(Vote), pitchers with more wins are seen as superior and should receive more votes.

Losses, is hypothesized to have a negative impact on Log(Vote), pitchers with a high number of losses are rarely noticed and are less likely to receive votes.

Saves, is hypothesized to have a positive impact on Log(Vote), saves mean a pitcher has earned a win for his team. Pitchers with a high number of saves should receive more votes.

IP, is hypothesized to have a positive impact on Log(Vote), pitching more innings benefits other pitchers on the team and signals positive performance. A pitcher with more innings pitched should receive more votes.

ERA, is hypothesized to have a negative impact on Log(Vote), a higher ERA means a pitcher allows more runs and should receive fewer votes.

K/IP, is hypothesized to have a positive impact on Log(Vote), higher strikeout rates makes it harder for other teams to score and should increase the number of votes the pitcher receives.

BB/IP, is hypothesized to have a negative impact on Log(Vote), walking a batter given the opposing team more chances to score and should decrease the number of votes the pitcher receives.

HR/IP, is hypothesized to have a negative impact on Log(Vote), pitchers who allow more home runs will have more runs scored against them which should decrease the number of votes received.

Starter, is hypothesized to have a positive relationship with Log(Vote), starting pitchers are more recognized than relievers and are expected to receive more votes.

V. Empirical Results

Dependent Variable: Log(Vote)
Method: Least Squares
Included observations: 298

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>-5.765838</td>
<td>0.993448</td>
<td>-5.803413</td>
<td>0.0000</td>
</tr>
<tr>
<td>Wns</td>
<td>0.241273</td>
<td>0.003276</td>
<td>7.452339</td>
<td>0.0000</td>
</tr>
<tr>
<td>Losses</td>
<td>-0.174008</td>
<td>0.003956</td>
<td>-5.104476</td>
<td>0.0000</td>
</tr>
<tr>
<td>Saves</td>
<td>0.077596</td>
<td>0.012775</td>
<td>6.083399</td>
<td>0.0000</td>
</tr>
<tr>
<td>IP</td>
<td>0.000507</td>
<td>0.000149</td>
<td>1.429505</td>
<td>0.1419</td>
</tr>
<tr>
<td>ERA</td>
<td>-0.512122</td>
<td>0.176642</td>
<td>-2.944361</td>
<td>0.0035</td>
</tr>
<tr>
<td>K/IP</td>
<td>-0.823940</td>
<td>0.444622</td>
<td>-1.878175</td>
<td>0.0622</td>
</tr>
<tr>
<td>BB/IP</td>
<td>-0.590589</td>
<td>0.728941</td>
<td>-0.831453</td>
<td>0.4076</td>
</tr>
<tr>
<td>HR/IP</td>
<td>2.444383</td>
<td>5.272200</td>
<td>0.539506</td>
<td>0.5900</td>
</tr>
<tr>
<td>Starter</td>
<td>-0.187791</td>
<td>0.943106</td>
<td>-0.199120</td>
<td>0.8423</td>
</tr>
<tr>
<td>Starter*K/IP</td>
<td>-2.371462</td>
<td>0.721260</td>
<td>-3.288167</td>
<td>0.0011</td>
</tr>
<tr>
<td>Starter*BB/IP</td>
<td>-0.738368</td>
<td>1.187789</td>
<td>-0.604190</td>
<td>0.5184</td>
</tr>
<tr>
<td>Starter*HR/IP</td>
<td>-7.547453</td>
<td>5.680996</td>
<td>-1.286184</td>
<td>0.2019</td>
</tr>
</tbody>
</table>

R-squared: 0.442779
Adjusted R-squared: 0.442779

VI. Conclusions

- Our adjusted R-squared indicates that 43% of the variation in percentage of the votes received is explained by our model.
- Wns are statistically significant in explaining the percentage of Cy Young votes received.
- Losses are statistically significant in explaining the percentage of Cy Young votes received.
- Saves are statistically significant in explaining the percentage of Cy Young votes received.
- ERA is statistically significant in explaining the percentage of Cy Young votes received.
- Starter*K/IP is statistically significant in explaining the percentage of Cy Young votes received.
- Our results used to create a predictive model for the Cy Young Award which correctly predicted 64% of the winners over the years in our data set. 88% of the time our model was able to place the winner in one of the top two vote getters for that given year.