**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.

**II. Empirical Model and Variables**

\[
\text{Log}(\text{Vote}) = \alpha + \beta_1 \text{Wins} + \beta_2 \text{Losses} + \beta_3 \text{Saves} + \beta_4 \text{IP} + \beta_5 \text{ERA} + \beta_6 \text{K/IP} + \beta_7 \text{BB/IP} + \beta_8 \text{HR/IP} + \beta_9 \text{Starter} + \beta_{10} \text{Starter*K/IP} + \beta_{11} \text{Starter*BB/IP} + \beta_{12} \text{Starter*HR/IP}.
\]

*Log(Vote)* is the 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 earned 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 allowed by the pitcher in that given season multiplied by the number of innings that they pitched.

**Starter** is dummy variable for whether the pitcher was a starter or reliever.

**III. Hypotheses**

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

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

*Saves* is hypothesized to have a positive effect on Log(Vote), as a pitcher with more saves should receive more votes.

*IP* is hypothesized to have a positive effect on Log(Vote), pitchers with more innings pitched are better at the game and signals positive performance. A pitcher with more innings pitched should receive more votes.

*ERA* is hypothesized to have a negative effect on Log(Vote), as a pitcher allows more runs and should receive fewer votes.

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

**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

**V. Empirical Results**

<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>-5.76389</td>
<td>0.99344</td>
<td>-5.80343</td>
<td>0.0000</td>
</tr>
<tr>
<td>Wins</td>
<td>0.24127</td>
<td>0.03237</td>
<td>7.452339</td>
<td>0.0000</td>
</tr>
<tr>
<td>Losses</td>
<td>-0.174008</td>
<td>0.035956</td>
<td>-5.124476</td>
<td>0.0000</td>
</tr>
<tr>
<td>Saves</td>
<td>0.077596</td>
<td>0.012775</td>
<td>6.083999</td>
<td>0.0000</td>
</tr>
<tr>
<td>IP</td>
<td>0.000507</td>
<td>0.000419</td>
<td>1.472905</td>
<td>0.1419</td>
</tr>
<tr>
<td>ERA</td>
<td>-0.511212</td>
<td>0.173624</td>
<td>-2.944361</td>
<td>0.0035</td>
</tr>
<tr>
<td>K/IP</td>
<td>-0.823940</td>
<td>0.644622</td>
<td>-1.278175</td>
<td>0.2222</td>
</tr>
<tr>
<td>BB/IP</td>
<td>-0.506598</td>
<td>0.739641</td>
<td>-0.707153</td>
<td>0.4794</td>
</tr>
<tr>
<td>HR/IP</td>
<td>2.844383</td>
<td>5.272200</td>
<td>0.539906</td>
<td>0.5900</td>
</tr>
<tr>
<td>Starter</td>
<td>-0.187791</td>
<td>0.934106</td>
<td>-0.199120</td>
<td>0.8423</td>
</tr>
<tr>
<td>Starter*K/IP</td>
<td>2.371623</td>
<td>0.721620</td>
<td>3.288167</td>
<td>0.0091</td>
</tr>
<tr>
<td>Starter*BB/IP</td>
<td>-0.738368</td>
<td>1.811778</td>
<td>-0.406190</td>
<td>0.6849</td>
</tr>
<tr>
<td>Starter*HR/IP</td>
<td>-7.547453</td>
<td>5.680896</td>
<td>-1.286184</td>
<td>0.1994</td>
</tr>
</tbody>
</table>

R-squared: 0.432307

R-squared (adj): 0.384207

F-test: 15.6701

Prob[F statistic]: 0.00000

**VI. Conclusions**

- Our adjusted R-squared indicates that 43% of the variation in percentage of the votes received is explained by our model.
- Wins 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 were 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.