Separating the Good From the Great: Predicting Votes for the Cy Young Award

Colin Nelson and John Henry

Linfield Department of Economics • Spring 2015

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 relieving.

II. Empirical Model and Variables

$$\text{Log (Vote)}_i = f(\text{Wins}_i, \text{Losses}_i, \text{Saves}_i, \text{IP}_i, \text{ERA}_i, \text{KIP}_i, \text{BBIP}_i, \text{HRIP}_i, \text{Starter}_i, \text{Starter*KIP}_i, \text{Starter*BBIP}_i, \text{Starter*HRIP}_i)$$

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

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

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

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

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

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

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

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

HRIP_i = Number of homeruns given up by the pitcher in that given season multiplied by the number of innings that they pitched.

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

*i denotes player where i = 1 - 268

III. Hypotheses

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

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

Saves_i is hypothesized to have a positive effect on Log(vote), saves mean a pitcher has ensured a win for his team. Pitchers with a high number of saves should receive more votes.

IP_i is hypothesized to have a positive effect 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_i is hypothesized to have a negative effect on Log(vote), a higher ERA means a pitcher allows more runs and should receive fewer votes.

KIP_i is hypothesized to have a positive effect on Log(vote), higher strikeout rates makes it harder for other teams to score and should increase the number of votes the pitcher receives.

BBIP_i is hypothesized to have a negative effect on Log(vote), walking a batter gives the opposing team more chances to score and should decrease the number of votes the pitcher receives.

HRIP_i is hypothesized to have a negative effect on Log(vote), pitchers who allow more homeruns will have more runs scored against them which should decrease the number of votes received.

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

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 Baseballrefence.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

Dependent Variable: LOG(VOTE)
Method: Least Squares
Included observations: 298

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-5.765389	0.993448	-5.803413	0.0000
Wins	0.241273	0.032376	7.452339	0.0000
Losses	-0.174008	0.033956	-5.124476	0.0000
Saves	0.077596	0.012755	6.083399	0.0000
IP	0.005037	0.003419	1.472905	0.1419
ERA	-0.511212	0.173624	-2.944361	0.0035
KIP	-0.823940	0.644622	-1.278175	0.2022
BBIP	-0.590589	1.589941	-0.371453	0.7106
HRIP	2.844383	5.272200	0.539506	0.5900
Starter	-0.187791	0.943106	-0.199120	0.8423
Starter*KIP	2.371625	0.721260	3.288167	0.0011
Starter*BBIP	-0.738368	1.817789	-0.406190	0.6849
Starter*HRIP	-7.547453	5.868096	-1.286184	0.1994

R-squared	0.453079	Mean dependent var	-3.064307
F-statistic	19.67493	Prob(F-statistic)	0.000000

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*KIP 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 as one of the top two vote getters for that given year.