Hall of Fame Voting Percentage Regression on Hitters Career Statistics

Kian Kurokawa

University of Hawaii Manoa

August 23, 2022

Abstract

The purpose of this project was to investigate the relationship between a players career batting statistic and his Hall of Fame voting percentage. To do this I ran a regression of Hall of Fame voting percentage on multiple explanatory variables including a players career Hits, career HR, and other career statistics. First, I collected career statistics on 493 MLB players that were on the Hall of Fame ballot. I excluded all pitchers to simplify the model and to achieve a consistent base for comparing hitters. Players on the Hall of Fame ballot before 1966 were also excluded due to the change of voting rules. After running the regression I found that there was a positive relationship between a players career statistics and Hall of Fame voting percentage. The results of the regression underestimate voting percentage based purely on a hitters career statistics suggesting that there may be a subjective "image" factor that plays a role in a players voting percentage.

1 Introduction

Baseball is often called "Americas Pastime" and for good reason. Its a game that has survived through both World Wars, player strikes, and even the intrusion of "stat geeks" that litter MLB front offices presently. Although Baseball is a constantly changing game, there are a few constants that have withstanded time. One constant is the structure of how baseball is played. Although Baseball is thought to be a team sport the majority of a Baseball game boils down to a single one on one interaction between the pitcher and batter. The outcomes of these countless interactions, giving us a large sample size, can be easily recorded and, as a result, it has been one of the best sports to do statistical analysis on.

Another constant has been the National Baseball Hall of Fame. It is the place where every kid dreams of making it to, the ultimate and final validation on a players outstanding career, and the difference from being remembered as a hero or living forever in the Halls as a legend. The inaugural National Baseball Hall of Fame class was reported to the media on February 2, 1936. Legends within this class included pioneers like Babe "The Great Bambino" Ruth and Walter "Train" Johnson. Fast forward 82 years and now the National Baseball Hall of Fame is now home to 323 elected members. Of the 323 elected member, 226 are former major league baseball players.

Despite the fact that the Hall of Fame itself has been constant, the process in which players are voted in has changed. Presently to be considered for the Hall of Fame a baseball player must have played in at least 10 major league baseball championship seasons. If a player makes this first cutoff, he must be elected by at least 2 baseball writers out of a committee of 6 to be eligible for the ballot. If a player has made it onto the voting ballot, he must receive at least 75% of the votes to get into the Hall of Fame. If a player fails to receive at least 5% of the votes for any year or has not been elected into the Hall of Fame within 10 years they are kicked off the ballot. The criteria

to earn the right to vote can be found in the index along with the changing history of the voting process.

On every Hall of Fame ballot, the criteria in which a Hall of Fame player is judged upon is "Voting shall be based upon the player's record, playing ability, integrity, sportsmanship, character, and contributions to the team(s) on which the player played." When I first read this criteria my initial thought was that this was immensely broad, especially if it is used for the Hall of Fame. I wanted to examine the relationship between a players career stats and how much percentage points they obtained while on the ballot. By examining this relationship I can conclude what career statistics are weighted more heavily than others as well as uncover certain subjective assessments made by the voting writers. As a result of examining this relationship I may uncover 'snubs' or players that should have made it into the Hall of Fame based on their career statistics which will raise the question of how well the current system of voting is performing.

2 Summary of Statistics Table

Table 1: Summary Statistics for Hall of Fame Ballot Hitters

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
vote	493	0.133	0.284	0.000	0.000	0.041	0.993 (Ken Griffey Jr.)
WAR	493	35.611	24.070	−3	19.5	46.4	163 (Barry Bonds)
HR	493	199.984	140.416	6	90	282	762 (Barry Bonds)
SB	493	123.260	149.473	0	31	156	1,406 (Rickey Henderson)
BA	493	0.273	0.021	0.193	0.260	0.287	0.344 (Ted Williams)
OPS	493	0.771	0.084	0.529	0.722	0.823	1.116 (Ted Williams)

My Summary Statistics table shows the mean, standard deviation, max, and min of each “traditional“ statistical category. I decided to use a majority of “traditional“ statistics in my regression since these were the only available statistics until the revolution of moneyball. There also seems to be discrete career milestones that a Hall of Famer must reach through the accumilataion of these “traditional” stats. For example, most voters often said that “A player who joins the 3000 hit club will most likely make it into the Hall of Fame“. However I did add the “modern“ statistics of OPS and WAR. OPS,which is just SLG and OBP summed together,better captures how impactful a hitter was adjusting for the different weights (bases) between singles, doubles, triples, and home runs. The statistic WAR, standing for Wins Above Replacement, calculates the total number of wins a player has gained for his team compared to an average player at his respective position.

Another decision I made was to exclude all pitchers, due to the fact that pitchers primarily focus on only pitching, as well as all players on the Hall of Fame ballot before 1966 due to different rule criteria for voting. These voting rule changes can be found in the Appendix.

3 Regression Results

Table 2: Results

	Dependent variable:	
	vote	
	(1)	(2)
H	0.0003***	-0.0001
	(0.00002)	(0.00004)
WAR		0.009***
		(0.001)
HR		0.001***
		(0.0002)
SB		-0.00001
		(0.0001)
BA		3.604***
		(1.003)
OPS		-1.530***
		(0.330)
Constant	-0.320***	-0.027
	(0.030)	(0.139)
Observations	493	493
R^2	0.346	0.526
Adjusted R^2	0.344	0.520
Residual Std. Error	0.230 (df = 491)	0.197 (df = 486)
F Statistic	259.502*** (df = 1; 491)	89.805*** (df = 6; 486)
Note:		*p<0.1; **p<0.05; ***p<0.01

Regression (1)

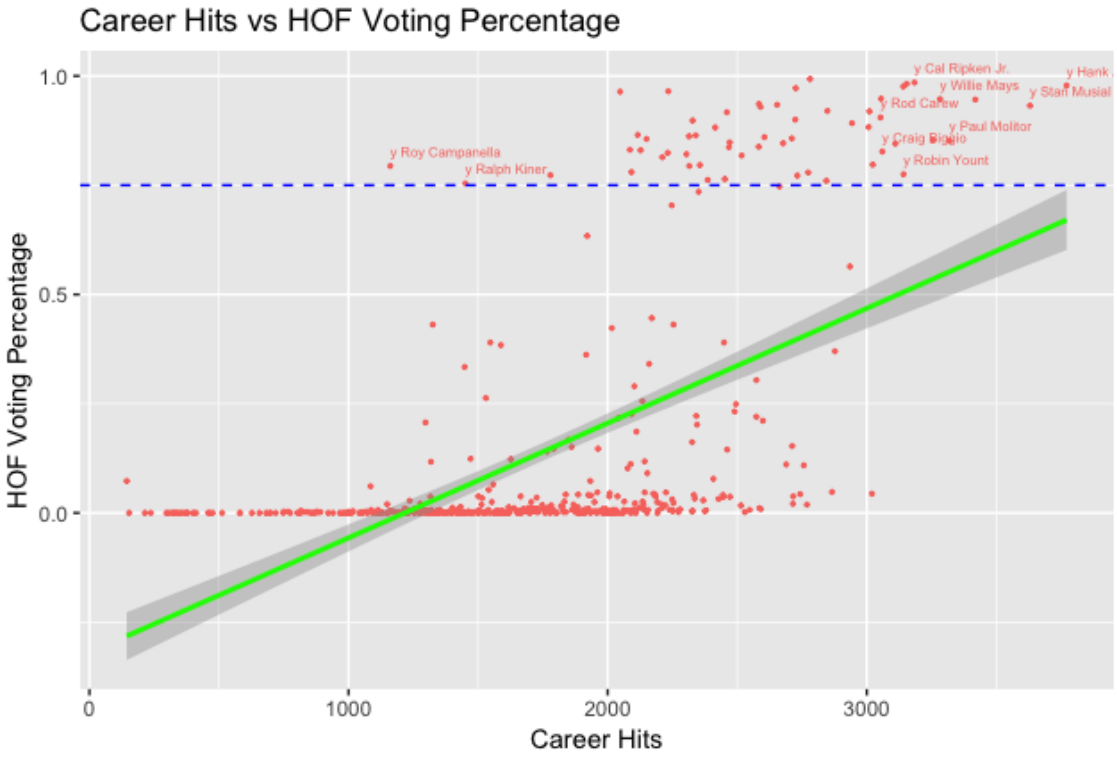

Figure 1: *Blue Dash indicates Cutoff Voting Percentage to Get into Hall of Fame

The first regression of voting percentage on hits reveals surprising results. It can be seen that all Hitters that enter the MLB start with -32% of getting to the Hall of Fame. Instead of hitters trying to perform to get up to 75%, the model suggests that they must obtain 107%! Although these are unexpected results, we know that a hitters career is an accumulation of other stats and factors which our regression does not consider. Even though we have statistical significance for the intercept and career hits, our R^2 and Adjusted R^2 tells us that there is a lot of information left in our observable term that can be used to better predict this relationship.

Knowing the shortcomings of the first regression, I ran a second regression of HOF voting per-

centages on a variety of hitting statistics. From our model we observe career WAR,HR,BA,OPS are statistically significant at the 99% confidence level. As a result of including these other statistics, career HITS is now statistically insignificant along with career SB and the intercept.

4 Classical Linear Model Assumptions

In order to "trust" our regression results, the six assumptions of a Classic Linear Model must be satisfied. If any assumption is not met we must adjust our model to satisfy each assumption.

4.1 Linear in Parameters

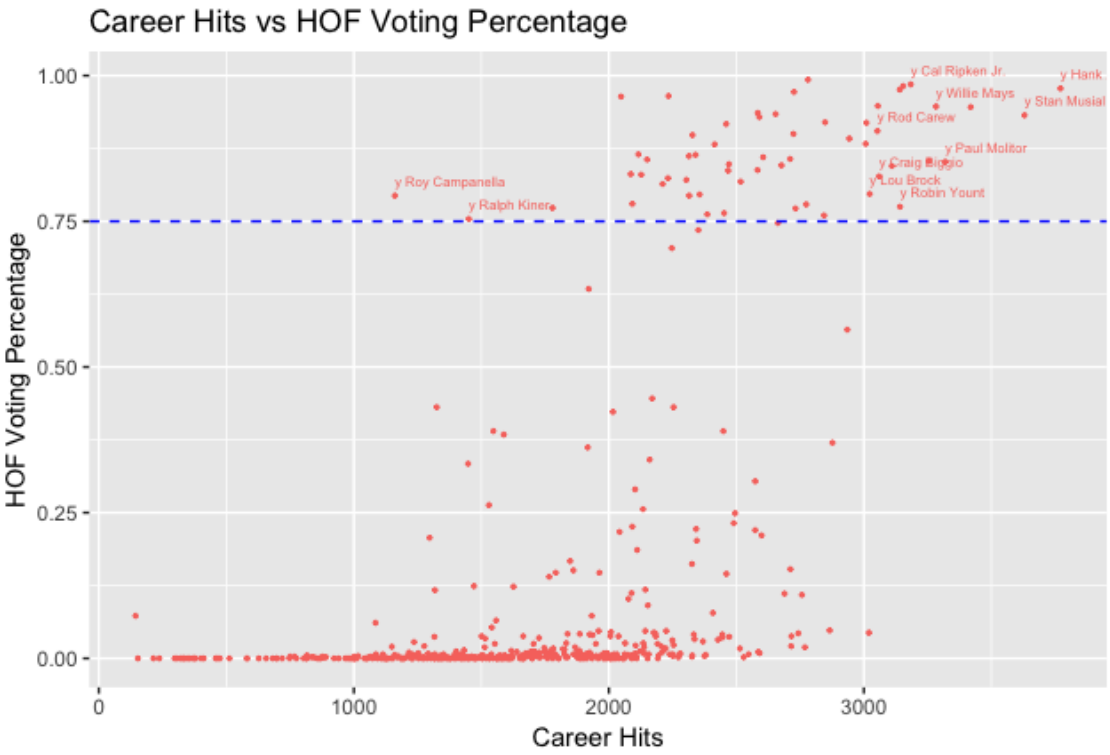

Figure 2

Assumption 1 states that we have a linear relationship between our dependent and independent variables. This assumption is met based on the context of the data. From figure 2, we can see that there is a linear relationship between accumulating higher career hitting totals and increasing HOF voting percentages.

4.2 Random Sampling

Assumption 2 states that the data is a random sample drawn from the population. I defined my population as all MLB players on the ballot starting from 1966 excluding pitchers. I then randomly drew players and their career stats to satisfy assumption 2.

4.3 Sample Variation in the Explanatory Variable

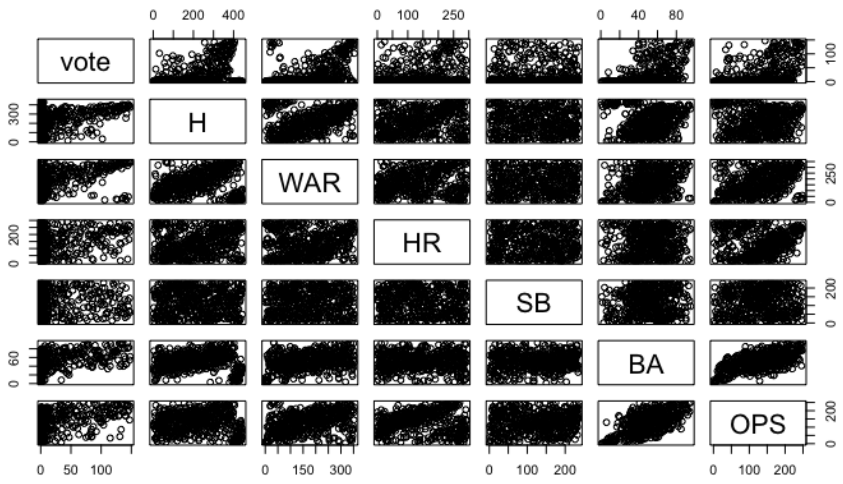

Figure 3

Assumption 3 states that there is no perfect co-linearity between our explanatory variables. Although there is no perfect co-linearity in our model, we must check for high levels of multicolinearity between our explanatory variables. To do this I produced scatter plots for each explanatory variable compared to the other explanatory variables as seen in figure 3. Although there is slight multicolinearity, especially between BA and OPS, there is enough variation between our explanatory variables to satisfy Assumption 3.

4.4 Zero Conditional Mean

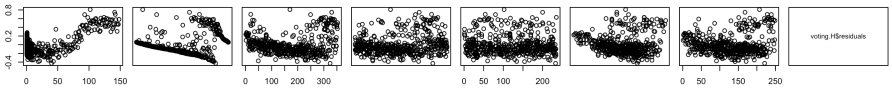

Figure 4

Assumption 4 states that the value of the explanatory variable must contain no information about the mean of the unobserved factors. To test this I created scatter plots of each explanatory variable with the residuals. Looking at each graph there seems to be no correlation between the explanatory variables and our unobserved factors. I also calculated the correlation between Hits,WAR,HR,SB,BA, and OPS with our unobserved factors and got 2.649416e-17, 2.649416e-17, 0.3077764, 0.1605059, 0.001641903, -0.008350084, 0.1215006 respectively. As a result of these correlations being small, assumption 4 is satisfied.

4.5 Homoskedasticity

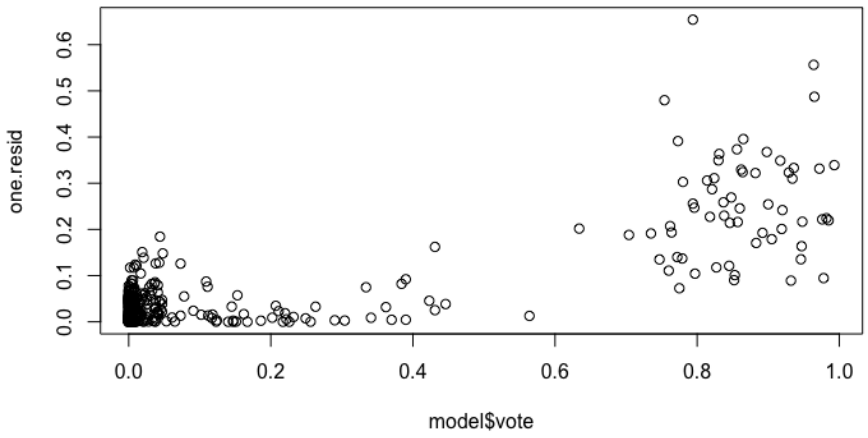

Figure 5

Assumption 5 states that the variance of our model should be constant. Refering back to figure 1 and figure 5, the squared residuals plotted against vote percentage, we can see that there is strong Heteroskedasticity. Running a Breusch-Pagan test on the model I obtained a p-value of 2.2e-16. With this value we can reject the null hypothesis of Homoskedasticity.

To achieve homoskedasticity I used the coeftest() command on r and obtained these results.

Table 3: Results

	Dependent variable:		
	vote		
	OLS		*coefficient test*
	(1)	(2)	(3)
H	0.0003***	-0.0001	-0.0001
	(0.00002)	(0.00004)	(0.00004)
WAR		0.009***	0.009***
		(0.001)	(0.001)
HR		0.001***	0.001***
		(0.0002)	(0.0002)
SB		-0.00001	-0.00001
		(0.0001)	(0.0001)
BA		3.604***	3.604***
		(1.003)	(0.997)
OPS		-1.530***	-1.530***
		(0.330)	(0.338)
Constant	-0.320***	-0.027	-0.027
	(0.030)	(0.139)	(0.148)
Observations	493	493	
R^2	0.346	0.526	
Adjusted R^2	0.344	0.520	
Residual Std. Error	0.230 (df = 491)	0.197 (df = 486)	
F Statistic	259.502*** (df = 1; 491)	89.805*** (df = 6; 486)	
Note:			*p<0.1; **p<0.05; ***p<0.01

4.6 Error Terms are Normally Distributed

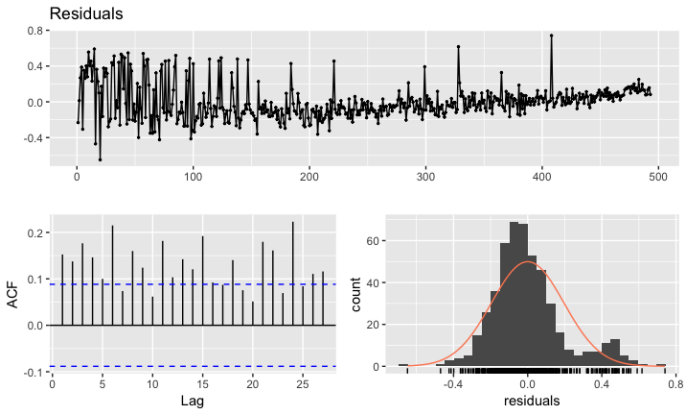

Figure 6

Assumption 6 states that the difference between our observed values and our fitted values are normally distributed. To check this I created an ACF plot, histogram, and line graph of our residuals. Looking at these graphs we can see that our residuals are normally distributed, satisfying assumption 6.

5 Conclusion

I have learned, based on our two regression models, that getting into the National Baseball Hall of Fame is difficult. With the thousands of former MLB players less than 1% that ever played make it into the Hall. This raises concern about the voting system and procedures of the Hall of Fame. As voters, their main job is to recognize Hall of Fame careers and filter out the others. However this is not always the case. Although my model undershoots the predicted voting percentage, this maybe due to the fact that "traditional" voting criteria were either 3000 career hits, 500 career

home runs, or a combination of the both, we notice certain "snubs" within our data.

5.1 Error Terms are Normally Distributed

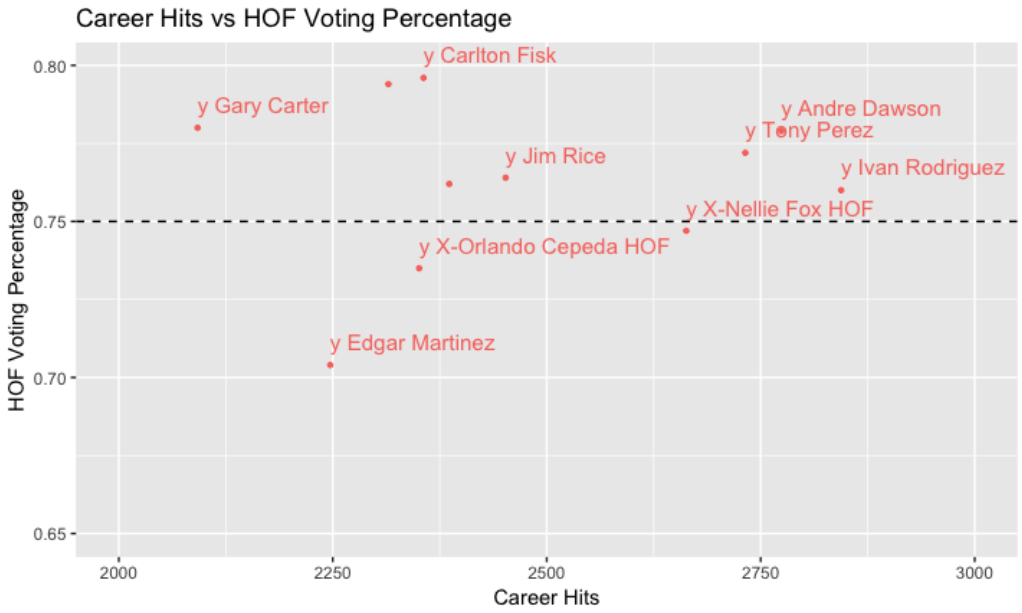

Figure 7

As seen in Figure 7, Edgar Martinez should be in the Hall of Fame if we are comparing him to Gary Carter. However, he is still approximately 5 percentage points away from being elected with his last year on the ballot coming up in 2019. This may be due to the fact that Edgar Martinez was primarily a Designated Hitter throughout his career. As a result, the voters perceived "image" of him has kept him out of the Hall of Fame for now, their main justification being that he didn't play defense. However based strictly on the data and comparison to other Hall of Fame players, he should already be in the Hall.

This raises the question of how we should evaluate a players career. Since being elected into

the Hall of Fame is a special honor, it is important to get every deserving players into the Hall. Hopefully in the future, voters will be more open to statistical analysis to give players like the Edgar Martinezs of baseball the recognition that they truly deserve.

Appendix

Link to find voting rules and changes: https://www.baseball-reference.com/about/hof$_voting.shtml$

R Code

Found as HTML file attached to document

References

[1]"Hall of Fame Ballot History." BR Bullpen, www.baseball-reference.com/awards/hall-of-fame-ballot-history.shtml.