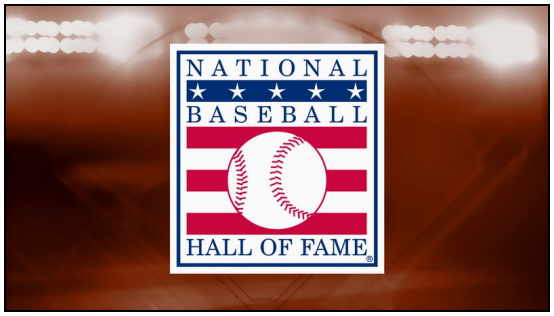


Predicting Baseball Hall of Fame Inductees

James Tang and Guy Dotan

UCLA Statistics 412 - December 13, 2019

Contents

1 Background

Located in Cooperstown, New York, Major League Baseball's Hall of Fame is an exclusive club filled with just 331 members representing over 150 years of professional baseball history. Members can be voted in for a variety of contributions to the sport and of those members, the breakdown is as follows: batters (178), pitchers (83), managers (23), umpires (10), and executives (35). The balloting process to be inducted is challenging by design as over 19,000 players have played the game professionally, but just around 1% of them have been inducted into the Hall of Fame.

In order to be inducted into Cooperstown, first, a player must have been retired for five years. After that time frame, there is a voting period each year where over 300 members of the BBWA (Baseball Writers' Association of America) and members of the Veterans Committee (certain retired members of the MLB) fill out a HOF ballot. In each ballot a voter can list up to 10 players that they believe deserve to be inducted. If a player appears on at least 75% of the ballots, then he is officially enshrined into this hallowed group of baseball legends.

The goal of this project was to build a model that could predict whether an upcoming nominee, based on his career resume, would be elected into the National Baseball Hall of Fame. To narrow our focus, we chose to only address pitchers.

2 Data Description

Major League Baseball, more than any other league, is known for its rich history intertwined with statistics and analytics. MLB's commitment to the importance of data-driven decisions in the team's front office opened the doors to the now thriving community of sports analytics. An example of which is our dataset: the famous Lahman Baseball Database, an open source repository of baseball records dating all the way back to 1871.

2.1 Variables Used

Because of the granularity of the Lahman baseball archives, there was much data preparation required. The raw dataset had season-by-season metrics for each pitcher. We aggregated these numbers in two different ways. First, we combined all the seasons together to get each player's career statistics (both regular season and postseason). The second aggregation method looked at the number of "high-achieving" seasons the player accumulated. This was defined as the count of seasons in which the pitcher ranked in the top 5 or 10 in one of four metrics: wins, strikeouts, ERA, and WAR.

WAR (Wins Above Replacement), mentioned earlier, is an advanced statistic developed and utilized by the field of Sabermetrics. Sabermetrics "aims to quantify baseball players' performances based on objective statistical measurements especially in opposition to many of the established statistics [...] that give less accurate approximations of individual efficacy."[1] Briefly, WAR provides a metric as to how many "wins" the particular player adds to their team compared to a league average player at that position. Not only is WAR a great metric because it is an all encompassing measure of a player's value, but it is also context agnostic. Essentially, WAR can be used to compare players across eras, leagues, ballparks, position, etc.

In addition to performance metrics, we also looked at awards and nominations for the pitchers. The relevant awards, which we totalled for the players career, included: All-Star selections, Most Valuable Player awards, Cy Young Awards (given to the league's best pitcher that season), and postseason MVP awards.

Finally, to determine our final list of pitchers to use in our dataset we limited the list of pitchers to those who had accumulated at least 1,000 innings pitched in their career. This ensured that our test cases had a robust resume to look into (all HOF pitchers reached this milestone). Additionally, we removed any pitchers whose entire career occurred before 1900. Generally, the 20th century is referred to as baseball's

[1] Encyclopedia Britannica: https://www.britannica.com/sports/sabermetrics

"modern era" so we decided to stick exclusively to players that fell within this distinction.

3.2 Feature Motivation

Including the career regular season, career postseason, season ranks, and awards statistics, we have around 70 total variables in our dataset. With a large variety of available features we have to consider a holistic understanding of both the statistics we include as well as the players they relate to.

There are two types of pitchers in baseball: starting pitchers and relief pitchers. Because starting pitchers begin the games and generally pitch more innings than relief pitchers, they usually accumulate much more counting stats (e.g. strikeouts, wins, earned runs). Saves are a statistic that credits relief pitchers, and since relief pitchers are in the hall of fame, it will be prudent to consider saves in our model.

When evaluating a player's resume it is important to emphasize both career longevity as well as career excellence. A Hall of Fame player needs to have made a sustained production over many years, but also some of those years need to have been impactful. Reaching certain career benchmarks (e.g. 100+ WAR, 300+ wins, 3000+ strikeouts) are major milestones on a resume. But just as important are single-season accolades like All-Star appearances, Cy Young Awards, and seasons ranking atop the league in a key statistic. Both types of features will need to be addressed in our model.

Finally, we would like to use a mix of traditional metrics as well as advanced ones. As mentioned earlier, WAR is a powerful tool to provide an apples-to-apples comparison across generations of the sport. That said, WAR was not a metric that was available until the last couple decades, and for the majority of the Hall of Fame history, voters did not have it at their disposal. They were forced to stick to standard statistics. Therefore, WAR is a useful gauge, but cannot be solely relied upon.

3 Exploratory Data Analysis

Before beginning any modeling steps we decided to run some exploratory analysis on our dataset to ensure there weren't any glaring irregularities. As a first step, we looked at boxplots of each of our variables in comparison to HOF induction (included in the Appendix). On a high level, it does appear that many of the counting stats (wins, hits, strikeouts, etc.) were associated with HOF induction. Rare events, like the awards variables, on the other hand, were not as clear of an association. Finally, WAR and our season-ranking metrics did seem correlated (except for save-related stats). The findings from these boxplots suggest further exploration into key metrics will be worthwhile.

This next stage of exploratory analysis involved a multivariable comparison of the traditional statistic (wins), our advanced stat (WAR), and their association with whether the player was inducted into the Hall of Fame. This chart is depicted in Figure 1. There were several key takeaways from this visualization. As we might expect, the majority of the clustering of Hall of Famers (green dots) are located in the top right of they chart—those with the highest career WAR and most career strikeouts. This suggests that both of these metrics are highly correlated with being inducted.

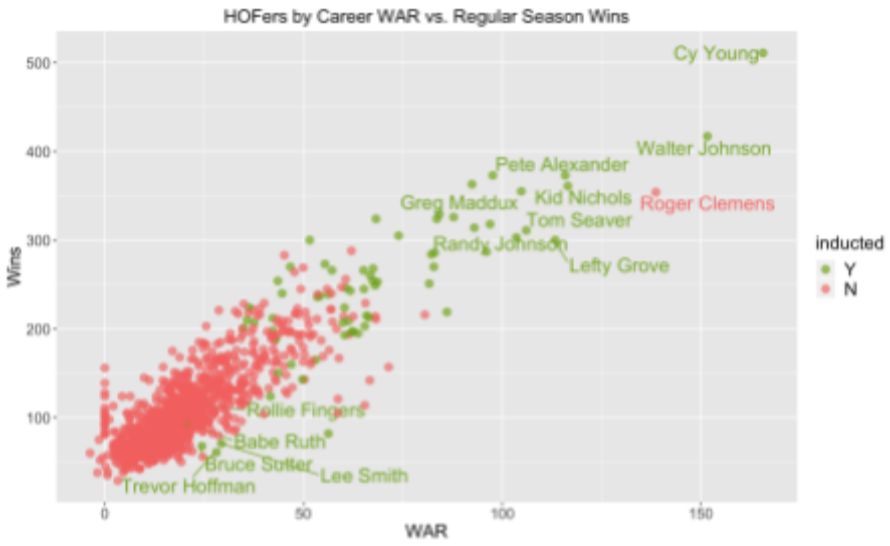


Figure 1 - War, Wins, and HOF

Notably, there is a major apparent outlier with one red dot at the top right: Roger Clemens. On the other side of the chart, in the bottom left, there are a few pitchers with very few wins and low career WAR that were inducted: Rollie Fingers, Trevor Hoffman, Lee Smith, and Bruce Sutter. All four of these pitchers were, in fact, relief pitchers and not starting pitchers. Finally, in that same bottom left corner is a very famous name, the most famous name in all of professional baseball: Babe Ruth. For the first six years of Babe Ruth's career he was a pitcher, and an extremely effective one at that. Upon get trading to the New York Yankees in 1920 he transitioned into a batter exclusively and put together potentially the best hitting resume of any player to this date. Although early in his career, Ruth was one of the best pitchers of his era, he was inducted into the HOF for his historical achievements at the plate, not on the mound.

Running this process again with new statistics yielded similar results. This time, comparing ERA, strikeouts, and HOF induction (Figure 2). Again we see a clustering of green dots in the top left region. ERA is a measure of, on average, how many earned runs a pitcher allows per full game. Thus the lower the mark, the better. As expected pitchers with lots of career strikeouts while giving up fewer runs per game was strongly correlated with getting inducted into the Hall of Fame. Just as before, however, the same clear outliers emerge: Roger Clemens (third most strikeouts ever) and Babe Ruth (only 488 career strikeouts).

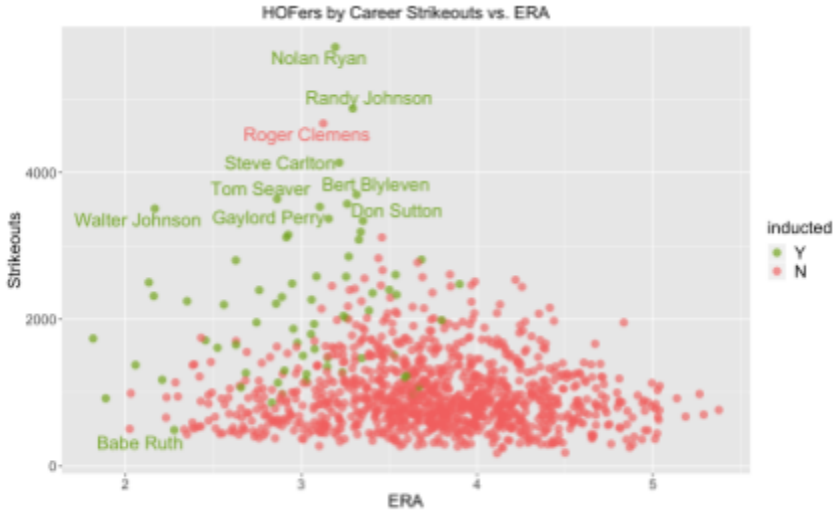


Figure 2 - Strikeouts, ERA, HOF

Takeaways from our exploratory analysis: (1) We should remove our outliers from the dataset. Babe Ruth made the Hall of Fame based on his hitting performance, not pitching. While Roger Clemens has not been inducted, despite his impressive resume, because of his implication in a significant steroid scandal. (2) Relief pitchers have a different looking resume than starting pitchers, and we will need to take that into account in our model.

4 Model Selection

4.1 Training Data and Test Data

In order to detect overfitting, we split our data into a training and test dataset. Our training set consisted of 95% of the non-HOFers and 70% of HOFers, and our test dataset consisted of 5% of the non-HOFers and 30% of the HOFers. We chose an asymmetric split in order to get a larger sample of HOFers into our test dataset.

4.2 Modeling Selection Strategy

The general framework of our modeling process was to use a binomial GLM and the step function for model selection. We configured the step function with direction=backwards and k=BIC, because these settings tended to be more stable (the step would often fail with other settings). Our primary goodness-of-fit metrics were accuracy on the test dataset and accuracy on HOFers (specificity) in the test dataset. For our confusion matrix, we chose a threshold of 50% to differentiate positive versus negative predictions. We also looked to minimize deviance, but solely minimizing deviance led to overfitting. We also examined the chi-square test (pchisq) for goodness-of-fit and for comparing nested models, but only using these tests to guide model selection also led to overfitting. In our investigation, our deviance numbers were significantly lower than our degrees of freedom in every model, and there were usually features we could add to significantly improve the model (according to the chi-squared test). In other words, underfitting was never an issue, so our discussion will focus on prediction accuracy on test data.

4.3 Initial Model

Our initial model (GLM using the step function and no interactions) gave us a 92.5% test-data accuracy and 75.0% specificity (accuracy on HOFers). Our relatively low specificity was not surprising given how imbalanced our dataset was - only 6% of the players in our dataset are in the Hall of Fame. Our model did not need to accurately predict HOFers to achieve a good overall accuracy. To counteract this, we decided to address this imbalance.

Variable	Coefficient
Intercept	-1.8935
Career wins	0.0338***
Career saves	0.0287***
Career ERA	-3.8902***
Top 5 wins	0.7363***
Top 10 strikeouts	0.3960**
Residual Deviance = 72.7	Degrees of Freedom = 944
· significant at p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

4.4 Weighting

To fix the imbalance, we weighted each HOFer in our training dataset by 8. We chose 8, because this gave us an even 30%/70% weighted-split for both our training set and data set. In our slides, we used oversampling, but we later decided to go with weighting (via svyglm from the survey package) as an alternative way to fix the imbalance. This approach gave us proper standard errors.

4.5 Model with Weights

Our model, using weighted training data and after removing insignificant features, improved accuracy (95.5%) and specificity (85.0%).

Variable	Coefficient
Intercept	-0.9753
Career wins	0.1624***
Career saves	0.0431***

Career ERA	-4.9437**
Career hits	-0.0095***
Career earned runs	0.0186*
Career walks	-0.0091***
Career strikeouts	-0.0040**
Top 10 strikeouts	0.9349***
Residual Deviance = 139.6	Degrees of Freedom = 941

˙ significant at p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Notice that the model with weighted data has a higher residual deviance (than our initial model: 139 vs 72), but has better accuracy. This is a sign that our initial model overfit the data.

4.6 Model with Weights and 2nd-Order Features

To improve our model, we added 2nd order features (interactions and square terms). We also added a transformation of ERA: 9 - ERA. We will refer to (9 - ERA) as "reverse ERA". This transformation converts ERA from a lower-is-better feature to a higher-is-better feature in hopes of creating a positive interaction with other higher-is-better stats (e.g. WAR).

The step function could not handle a model with all 2nd order terms, due to various issues (memory, runtime, convergence, etc), so we coded our own step function.

Our model selection algorithm was:

1. Among all possible features, square features, and interaction features, pick the one which lowers deviance the most.
2. Repeat until accuracy (on the test dataset) stops improving.

The resulting model with interaction terms improved our accuracy (98.5%) and specificity (95.0%). Because the significance level (0.06) of WAR was relatively close to the usual 0.05 cutoff, we decided to include it in our final model.

Variable	**Coefficient**
Intercept	-12.0566***

career WAR : career 9-ERA	0.0106•
career Saves : all star games	0.0052***
top 5 wins : career 9-ERA	0.1308***
career wins : career 9-ERA	0.0053**
Residual Deviance = 171.7	Degrees of Freedom = 945

• significant at p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001,

The interpretation of the coefficient for the interaction of career WAR and career reverse ERA is: A one unit increase in career WAR x career reverse ERA increases the odds of being inducted into the Hall of Fame by 1.07%. Note: exp(0.0106) = 1.0107. The other coefficients are interpreted in a similar way. We can see that our final model contains features which reward a mix of:

1. career longevity (career WAR, career wins, career saves)
2. career excellence (career ERA)
3. single season excellence (all star games, top 5 wins)

It is not enough for a pitcher to pitch for a long time and accumulate counting stats. Our model indicates that voters are also looking for periods of excellence. We also see a mix of traditional stats (wins, ERA) and advanced stats (WAR), and a mix of starting pitcher stats (wins) and relief pitcher stats (saves). This mix of stats helps address the blindspots of any feature by itself. Saves credits relief pitchers, which counting stats tend to underestimate. WAR captures performance contributions missed by traditional stats, but traditional stats were the primary metrics that HOF voters had access to (since WAR is a recent invention).

4.7 Model Diagnosis

According to the chi-squared goodness-of-fit test with residual deviance 171.1 on 945 degrees of freedom, our final model fits the training data well. Our final model's confusion matrix on test data is:

	Actual - No	**Actual - Yes**
Predicted - No	47	1
Predicted - Yes	0	19

Our final model's accuracy on test data is 98.5% and the specificity is 95.0%. For comparisons, the null model's accuracy is 70.1% and its specificity is 0%.

In the residual vs linear prediction plot (see Appendix 7.2), we do not see an even variation in the residuals along the linear predictor values, indicating that there is an inadequacy in the model. This is a topic of future research.

The half normal plot (see Appendix 7.2) reveals one outlying point, which corresponds to Bob Shawkey, who is currently not in the Hall of Fame. There is a lot of debate surrounding his omission from the HOF. We decided to leave Shawkey in the dataset, because his omission reflects the natural uncertainty in HOF voting.

4.8 Neural Network

As a last modeling attempt, we decided to take a look at another technique besides regression: a neural network. Because our dataset contains a lot of input variables and a very distinct, binary response, it seemed like a suitable problem for a neural network to learn. In order to build out this model we took all of our metrics and standardized them so they all had mean 0 and a standard deviation of 1. Because all our inputs are numerical variables this was achievable. Some variables had to be removed because we did not have that particular stat for that player and replacing the NULL values with 0 might skew our results.

For the sake of brevity, we did not apply any oversampling or cross-validation, but we did split the dataset into a 70:30 train and test set. Because of the limitations we had with computing power we used a single layer neural network and did not place all the variables into the model. Instead, we used a mix of our traditional, advanced, and starting and relief pitcher focused, and award-based statistics. Our final neural network contained 27 inputs, 30 nodes in the hidden layer, and a binary response classification. In the end, our model returned very promising results as the network built on our training set yielding 96.8% accuracy on the test set. With such promising results from the neural network, it appears that it would be worthwhile when conducting further research to build a deeper neural network with more hidden layers. In addition, a larger

network as well as a way to transform the variables that had missing values, would allow us to include all of these metrics into a more powerful neural network.

5 Predictions on Upcoming Nominees

As a method to apply and further analyze our prediction power we decided to input some active and recently retired players into our model. This would allow us to get a sense of the chances the player would be inducted into the Hall of Fame based on their career resume up to this point. Table 1 showcases our predictions for the players with ten highest odds of making the Hall of Fame.

Player	HOF Pred% (95% CI)	Player	HOF Pred% (95% CI)
Tim Hudson	66.2% (57.8%-73.6%)	Zack Greinke	9.3% (3.8%-20.9%)
Justin Verlander	57.2% (30.6%-80.1%)	Adam Wainwright	6.4% (3.3%-12.0%)
Mark Buehrle	31.2% (22.8%-41.1%)	Felix Hernandez	5.2% (2.8%-9.7%)
Clayton Kershaw	28.7% (9.5%-60.8%)	Jered Weaver	2.4% (1.2%-4.8%)
Bartolo Colon	26.7% (18.2%-37.5%)	Max Scherzer	1.7% (0.5%-4.9%)

Table 1: Final model predictions for active players

Some players are have pretty large confidence intervals (e.g. Kershaw and Verlander). Kershaw and Verlander had stellar career numbers, but had only played 8 and 10 seasons, respectively, and we can assume by the end of their careers their entire resume will put them more in line with the other Hall of Fame pitchers.

Compared to our generalized linear model, the neural network was a bit more generous for predicting current Hall of Famers, suggesting eight future inductees would make the cut (based on a threshold of 50% odds or greater). Table 2 shows our neural network predictions for the players with 10 highest odds of making the Hall of Fame (confidence intervals were not available).

Player	HOF Pred%	Player	HOF Pred%
Bartolo Colon	100.0%	Max Scherzer	96.8%
Tim Hudson	100.0%	Clayton Kershaw	77.0%

Justin Verlander	100.0%	CC Sabathia	63.7%
C. J. Wilson	100.0%	Adam Wainwright	5.1%
Felix Hernandez	99.7%	James Shields	2.0%

Table 2: Neural network predictions for active players

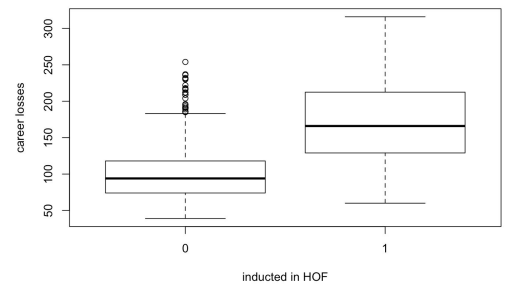
6 Further Research

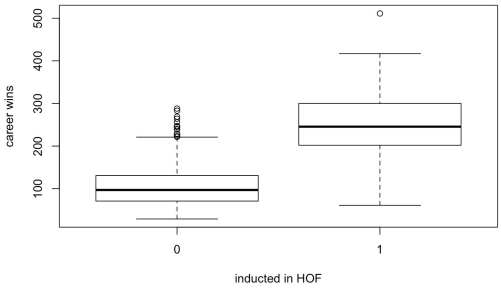
Our report explored some approaches to building a hall of fame prediction model using the given dataset. That said, there is potential to improve upon our methods. First, there is a world of other Sabermetrics statistics that could be utilized to increase predictive powers. These metrics are designed to define, previously indefinable qualities of baseball players and these could prove invaluable for model building. We believe it would be worthwhile to look into more of those in future modeling iterations. Secondly, we could look deeper into differences in the historical eras of players. Baseball is an ever-evolving sport and there could definitely be trends and shifts in these eras that could be baked into a better model. Lastly, an extremely powerful, but potentially overly trivial, variable to include would be Hall of Fame voting. Every year a player is eligible for induction the writers can give them votes. A trend in the players voting, either increasing in appearances on ballots or decreasing, would be a strong indicator of whether the player's chances of making the Hall of Fame are improving.

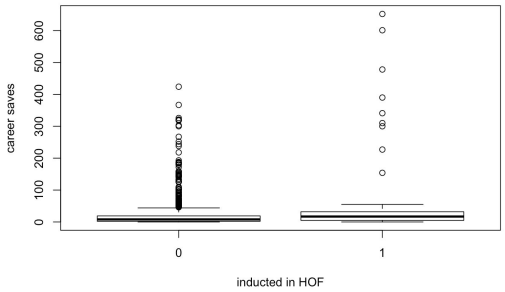
Additionally, for modeling clarity, we only addressed for pitchers and their statistics for this project. Doing the same for hitters would be a similarly challenging endeavor and might yield entirely different conclusions regarding the importance of variable types. That is: counting stats, advanced metrics, seasonal excellence, awards, and so forth. A hitters dataset would also incur more instances of steroid-related outliers, as an entire era of hitters would be subject to this controversy. That said, there are over twice as many batters who made the Hall of Fame than pitchers, so that could help model building and performance.

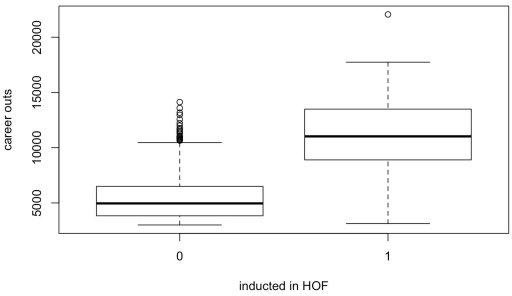
7 Appendix

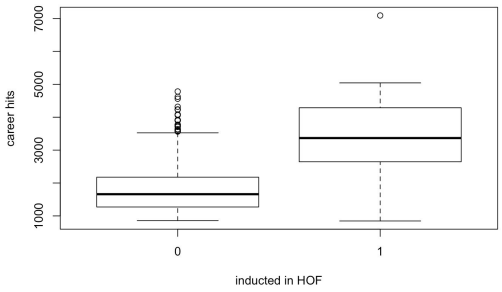
7.1 Explanatory vs. Response Variable Boxplots

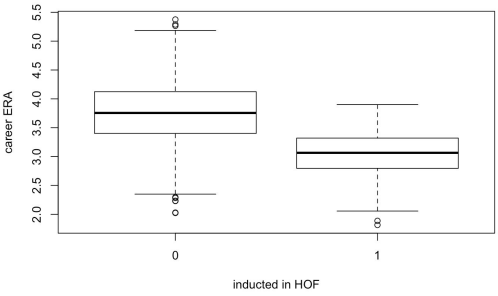


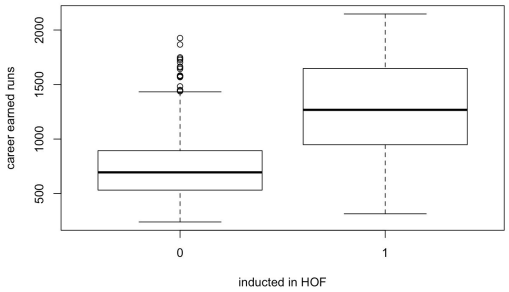


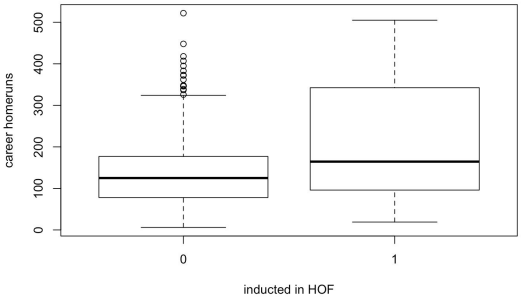


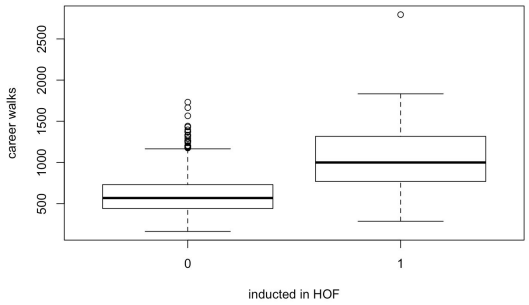


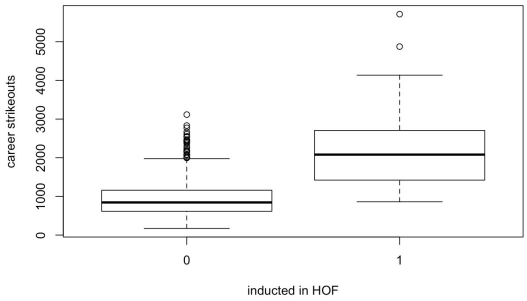


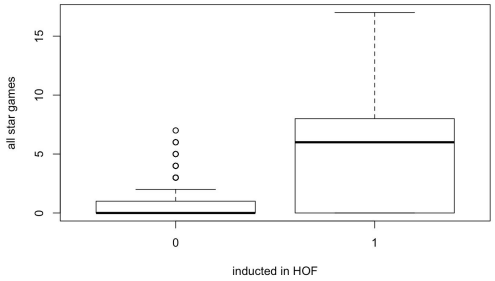


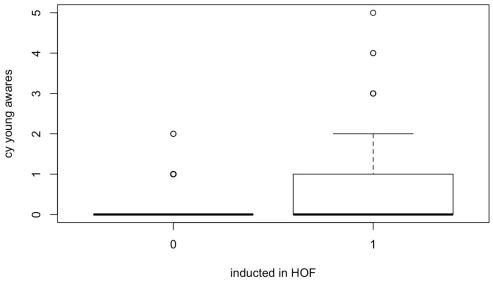


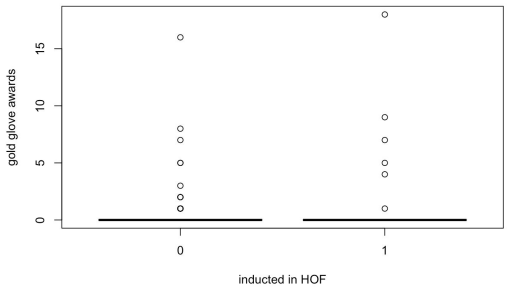


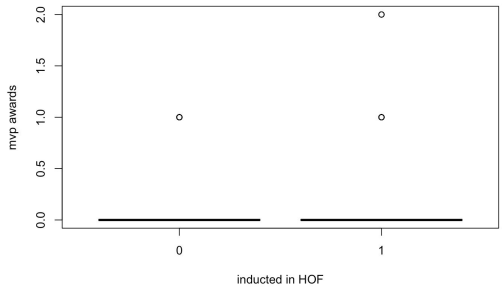

career walks
500
1000
1500
2000
2500
0
1
inducted in HOF

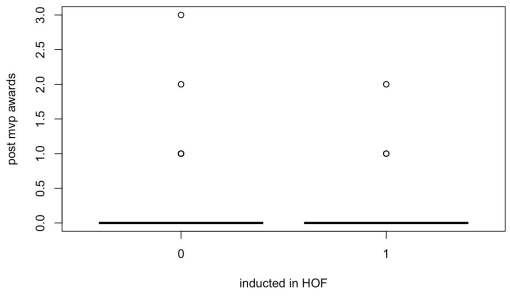

career strikeouts
0
1000
2000
3000
4000
5000
0
1
inducted in HOF

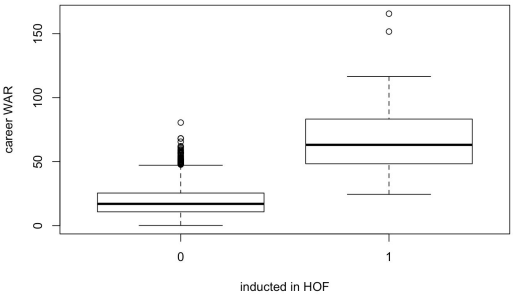

all star games
0
5
10
15
0
1
inducted in HOF

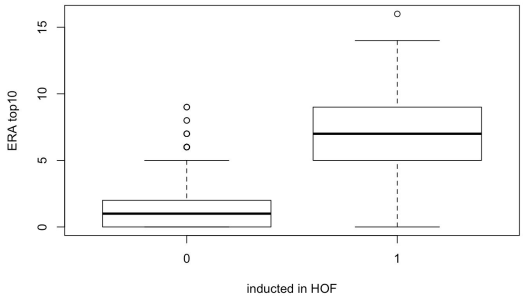

cy young awares
0
1
2
3
4
5
0
1
inducted in HOF

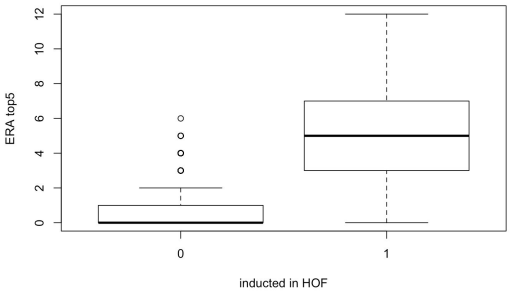

gold glove awards
0
5
10
15
0
1
inducted in HOF

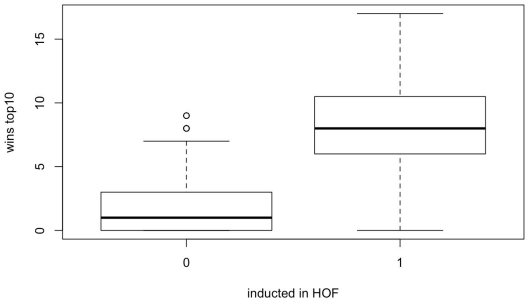

mvp awards
0.0
0.5
1.0
1.5
2.0
0
1
inducted in HOF

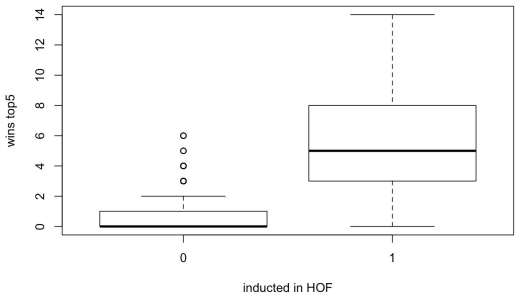

post mvp awards
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
inducted in HOF

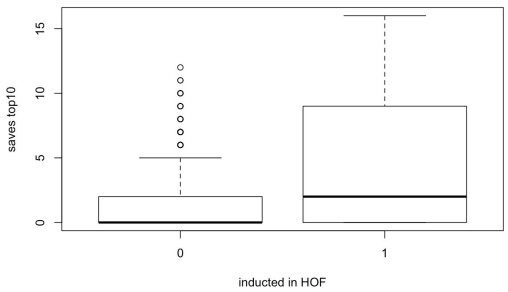

career WAR
0
50
100
150
0
1
inducted in HOF

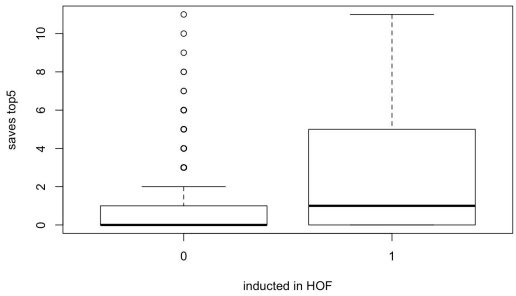

ERA top10
0
5
10
15
0
1
inducted in HOF

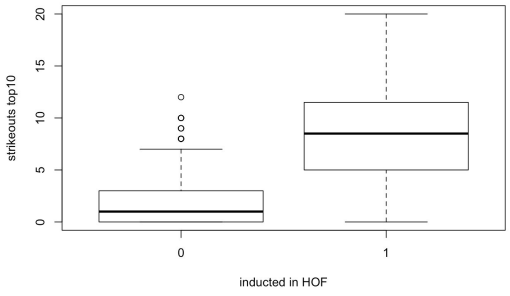

ERA top5
0
2
4
6
8
10
12
0
1
inducted in HOF

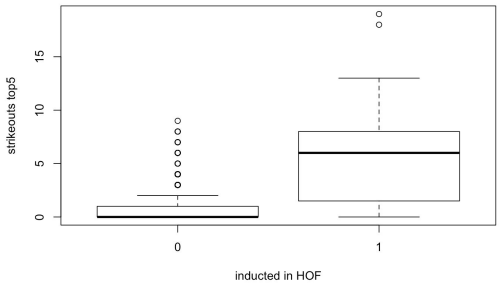

wins top10
0
5
10
15
0
1
inducted in HOF

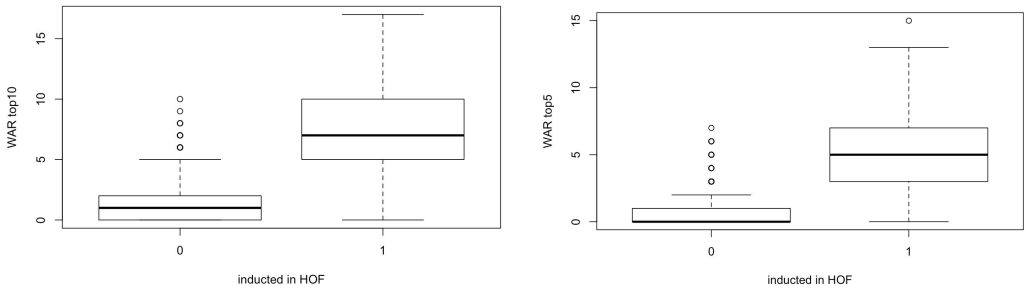

wins top5
0
2
4
6
8
10
12
14
0
1
inducted in HOF


saves top10
0
5
10
15
0
1
inducted in HOF

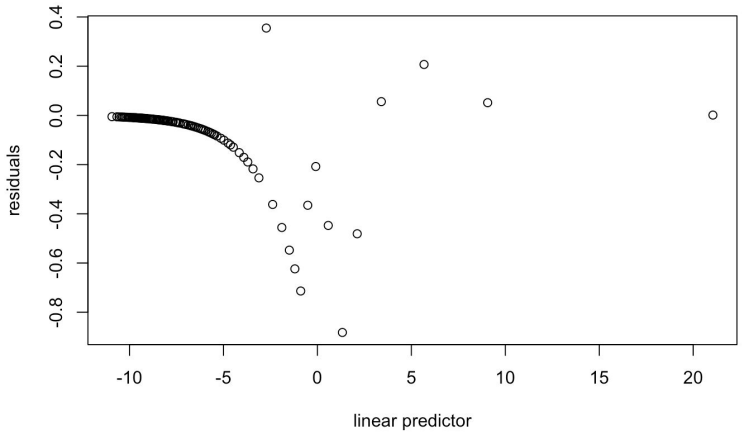

saves top5
0
2
4
6
8
10
0
1
inducted in HOF

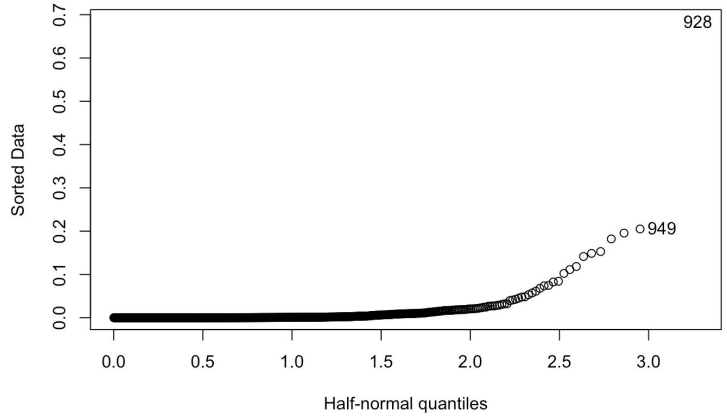

strikeouts top10
0
5
10
15
20
0
1
inducted in HOF


strikeouts top5
0
5
10
15
0
1
inducted in HOF



7.2 Final Model Residual Plots





8 References

Baseball's Modern Era:
https://www.history.com/news/what-is-baseballs-modern-era

Faraway, Julian James. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC, 2016.

Lahman Baseball Database:
http://www.seanlahman.com/baseball-archive/statistics/

National Baseball Hall of Fame:
https://baseballhall.org/hall-of-famers

WAR Dataset:
https://www.beyondtheboxscore.com/2013/2/7/3962494/saberizing-a-mac-8-creating-a-war-database