Baseball Hall of Fame Report

Andy, Dante, Kai, and Kobe

2023-03-03

Introduction

The data we are working with is career baseball statistics. The data was collected from official scorekeepers who went to every Major League Baseball game and tracked various outcomes players had over the course of each game. For our finished data set, those statistics were added up at the end of each baseball season and also at the end of each player's career. The motivation for the scorekeepers to track the data was that they were paid, either by the home team, the league as a whole, or the local newspaper to track these statistics to keep baseball fans informed.

One motivation for our research question was to better understand which variables are associated with a successful hall of fame induction. Another motivation for our research is an extension of the first, which is for us to predict which players will be inducted into the hall of fame in the future.

As such, our first research question is what factors are associated with a position player being inducted into the hall of fame and whether we can create a model that uses these factors to predict what players will make the hall of fame. Our other research question is what factors are a associated with whether a pitcher is inducted into the hall of fame and whether we can create a model that uses these factors to predict what pitchers will make the hall of fame.

Data Description

Our data consists of mostly quantitative variables mixed in with some categorical variables. Most baseball metrics come in quantitative form so it is often hard to obtain categorical measurements unless they come directly from the quantitative variables itself. However, some categorical variables we may use include factors such as year, league, team, stint, and possibly some other categorical variables that are computed from other quantitative variables. One other potential categorical variable would be a binary variable indicating whether a player was known to use steroids or not.

Furthermore, we are splitting this project into two parts: batters and pitchers. Thus, there is a need to split the data; in other words, we'll have two different data sets, since metrics for batters and pitchers are vastly different. However, we'll still be predicting the same variable: hall of fame. Hall of fame is a binary categorical variable (with the two values being 1 and 0), indicating whether a player was inducted into the hall of fame or not. In addition, we made sure to only include players who were on a hall of fame ballot at some point. In total, we have around 750 observations in the position players training data set and 400 observations in the pitching training data set.

Exploratory Data Analysis

Steroids Clustering

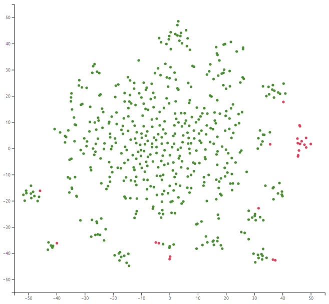


Figure 1: Clustering with Steroids in Red

Methods

Each of our pitcher and batter (position player) data frames consisted of 50 base variables after the conclusion of our various join functions.

For batters, we began with the "Batting" data frames within the Lahman package that included recognizable rate and counting stats for each observation (player) like batting average, walks, plate appearances, and total hits for each of the four base categories. We then joined a

For pitcher,

In addition to the base covariates, we saw a necessity to mutate additional variables

For our model building, we used essentially every possible variable that could potentially be useful and did not contain too many missing values. As such, we started our process with even larger models than ones we already had To narrow down the number of covariates, we used step-wise model selection that goes both forwards and backwards, allowing us to test as many models as possible. Since we are more focused on the prediction aspect, we chose to use AIC as our primary criteria for model selection over other criteria such as BIC or K-fold CV. In addition, AIC uses less computational power than LOOCV, which is vital for us given the number of covariates we used in building our models.

Our final models are as follows:

- Position Players:

  `wein ~ all_star + Pos + HR + nice_guy_awards + Steroids + votedBy + RBI + AVG + most_valuable_player + R + SB + Pos:HR + all_star:most_valuable_player + Steroids:SLG`

- Pitchers:

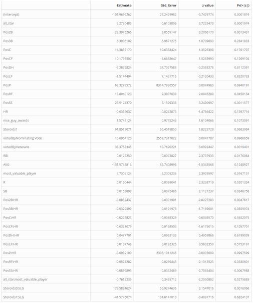
  `wein ~ W + L + SO + SV + Steroids + most_valuable_player + all_star + cy_young_award + rolaids_relief_man_award + votedBy + StarterReliever + Steroids:most_valuable_player`

Results

In terms of our model's performance on diagnostics on the training model, we will examine the coefficients and qqplots

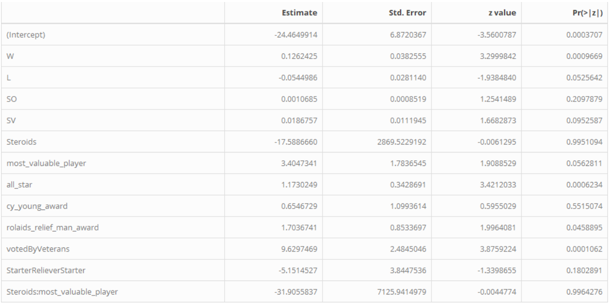
Coefficients

Position Players Coefficients



	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-101.9699262	27.2429982	-3.7429774	0.0001819
all_star	2.2720485	0.6103806	3.7223473	0.0001974
Pos2B	28.3975266	8.8559147	3.2066170	0.0013431
Pos3B	6.3906102	5.9671275	1.0709693	0.2841833
PosC	14.3832170	10.6334424	1.3526398	0.1761707
PosCF	10.1793507	6.6688647	1.5263993	0.1269104
PosDH	-8.2879824	34.7027588	-0.2388278	0.8112391
PosLF	-1.5144494	7.1421715	-0.2120433	0.8320733
PosP	62.3279572	8314.7926557	0.0074960	0.9940191
PosRF	18.8040120	9.3807638	2.0045289	0.0450134
PosSS	26.5124379	8.1599336	3.2490997	0.0011577
HR	-0.0358637	0.0242873	-1.4766422	0.1397716
nice_guy_awards	1.5742124	0.9775248	1.6104066	0.1073091
Steroids1	91.8512071	50.4019850	1.8223728	0.0683984
votedByNominating Vote	10.6964120	2559.7317022	0.0041787	0.9966659
votedByVeterans	33.3758345	10.7690221	3.0992447	0.0019401
RBI	0.0175250	0.0073827	2.3737935	0.0176064
AVG	-131.5742813	85.7408996	-1.5345568	0.1248927
most_valuable_player	7.7305124	3.2305235	2.3929597	0.0167131
R	0.0160444	0.0069041	2.3238719	0.0201324
SB	0.0153099	0.0072486	2.1121237	0.0346758
Pos2B:HR	-0.0852437	0.0301991	-2.8227283	0.0047617
Pos3B:HR	-0.0329599	0.0191973	-1.7169001	0.0859974
PosC:HR	-0.0222823	0.0368329	-0.6049570	0.5452075
PosCF:HR	-0.0321079	0.0198503	-1.6175015	0.1057701
PosDH:HR	0.0477701	0.0963133	0.4959866	0.6199039
PosLF:HR	0.0107748	0.0192326	0.5602350	0.5753191
PosP:HR	-0.6939100	2306.1011245	-0.0003009	0.9997599
PosRF:HR	-0.0574282	0.0269445	-2.1313525	0.0330601
PosSS:HR	-0.0899895	0.0332489	-2.7065404	0.0067988
all_star:most_valuable_player	-0.7613239	0.3455712	-2.2030882	0.0275885
Steroids0:SLG	179.5891624	56.9274636	3.1547016	0.0016066
Steroids1:SLG	-41.5776074	101.6141010	-0.4091716	0.6824137

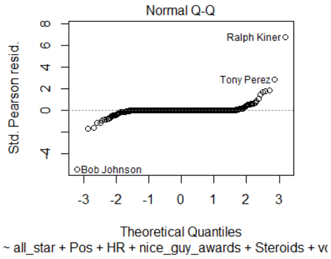
Pitchers Coefficients



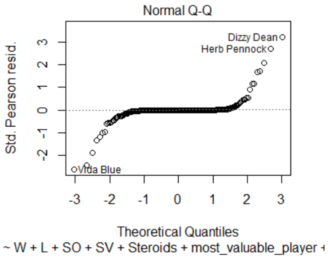
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-24.4649914	6.8720367	-3.5600787	0.0003707
W	0.1262425	0.0382555	3.2999842	0.0009669
L	-0.0544986	0.0281140	-1.9384840	0.0525642
SO	0.0010685	0.0008519	1.2541489	0.2097879
SV	0.0186757	0.0111945	1.6682873	0.0952587
Steroids	-17.5886660	2869.5229192	-0.0061295	0.9951094
most_valuable_player	3.4047341	1.7836545	1.9088529	0.0562811
all_star	1.1730249	0.3428691	3.4212033	0.0006234
cy_young_award	0.6546729	1.0993614	0.5955029	0.5515074
rolaids_relief_man_award	1.7036741	0.8533697	1.9964081	0.0458895
votedByVeterans	9.6297469	2.4845046	3.8759224	0.0001062
StarterRelieverStarter	-5.1514527	3.8447536	-1.3398655	0.1802891
Steroids:most_valuable_player	-31.9055837	7125.9414979	-0.0044774	0.9964276

QQ-Plots:

Position Players Coefficients



Pitchers Coefficients



As we can tell, for both the hitting and pitching models, the models have a similar fit to the training data.

Prediction Our final prediction results are published on the web application here