Major League Baseball Playoff Appearance Likelihood

Shannon Leiss

12/7/2021

Introduction

Many different attempts have been made to optimize an equation for predicting a team's probability of making the playoffs in all sports.

`Teams` dataset in the `Lahman` package provides team records for the 1871 - 2020 baseball season, along with offensive statistics, defensive statistics, and if the team made the playoffs.

Using this data for the seasons of 1998 - 2019, offensive statistics will be used to model the likelihood of a team making the playoffs.

Based off the best models, the offensive statistics needed to have a 50% and 90% chance of making the playoffs will be found.

Using the same offensive data, it will be compared which league - American or National - hits more home runs and which franchises historically hit the most home runs per season.

Offensive Statistics Being Considered

The offensive statistics used initial were:

- Runs Scored
- Hits
- At Bats
- Doubles
- Triples
- Homeruns

Addition variable that were calculated for use in a separate model were:

- Run Average: Runs Scored divided by At Bats
- Batting Average: Hits divided by At Bats

Data Used:

	ANA	ARI	ATL	BAL	BOS
R	787.0000	665.0000	826.0000	817.0000	876.0000
AB	5630.0000	5491.0000	5484.0000	5565.0000	5601.0000
H	1530.0000	1353.0000	1489.0000	1520.0000	1568.0000
X2B	314.0000	235.0000	297.0000	303.0000	338.0000
X3B	27.0000	46.0000	26.0000	11.0000	35.0000
HR	147.0000	159.0000	215.0000	214.0000	205.0000
RperAB	0.1398	0.1211	0.1506	0.1468	0.1564
HperAB	0.2718	0.2464	0.2715	0.2731	0.2800
Playoffs	0.0000	0.0000	1.0000	0.0000	1.0000

Initial Logistic Model

Since the response variable of `Playoff` is a indicator of whether or not a team made the playoffs, logistic regression can be used to model the log-odds of a team making the playoffs. The initial logistic model that was fitted:

$$log(\frac{p_i}{1-p_i}) = 25.306 + 0.0178(R_i) - 0.0068(AB_i) - 0.00015(H_i) \\ - 0.0035(X2B_i) - 0.0112(X3B_i) - 0.003(HR_i)$$

Where p_i is the probability that a team made the playoffs.

This model had an AIC of 673.9, a residual deviance of 659.9 on 653 degrees of freedom, with no evidence of over dispersion.

A model with all possible interactions was also fitted and was found to be worse than the initial model with no interactions.

Reduced Models

From the initial model, only the intercept, Runs Scored, and At Bats were significant. The reduced models that were fit:

$$log(\frac{p_i}{1-p_i}) = 26.2955 + 0.0162(R_i) - 0.0071(AB_i)$$

$$\begin{aligned} log(\frac{p_i}{1-p_i}) = &-115.292 + 0.1980(R_i) \\ &- 0.01837(AB_i) - 0.00003(R_i * AB_i) \end{aligned}$$

The AIC value for the model without interactions was 667.51 and the model with interactions had an AIC of 664.75.

Comparing Initial and Reduced Models

Using drop in deviance tests with likelihood ratios(no over dispersion), the difference in deviance(T) was compared to a χ^2_d distribution, where d is the difference in terms between models.

Comparing the initial model to the reduced model without interactions:

- T $=$ 1.6161 with a p-value $=$ 0.8059
- **Reduced Model** should be used

Comparing the reduced model with interactions to the model without interactions:

- T $=$ 4.763 with a p-value $=$ 0.02908
- **Model with Interactions** should be used

Logistic Model using Mutated Offensive Statistics

Using the created variables of Run Average and Batting Average per team, the following logistic model with interaction was fit:

$$log(\frac{p_i}{1-p_i}) = -41.92 + 330.99(RperAB_i) + 110.21(HperAB_i) - 922.93(RperAB_i * HperAB_i)$$

This model had an AIC of 674.71 and a residual deviance of 666.71 on 656 degrees of freedom.

Reduced Model without interaction:

$$log(\frac{p_i}{1-p_i}) = -7.808 + 87.704(RperAB_i) - 19.536(HperAB_i)$$

With an AIC of 675.32 and a residual deviance of 669.32 on 657 degrees of freedom.

Initial vs Reduced Model

Using drop in deviance test between above models:

- T = 2.6124, with p-value = 0.106
- **Model without Interaction** should be used

Optimal Models

Using Offensive statistics:

$$(1) \qquad log(\frac{p_i}{1-p_i}) = -115.292 + 0.1980(R_i) - 0.01837(AB_i) - 0.00003(R_i * AB_i)$$

Using Mutated Offensive statistics:

$$(2) \qquad log(\frac{p_i}{1-p_i}) = -7.808 + 87.704(RperAB_i) - 19.536(HperAB_i)$$

Different Link Functions on Models

Using the two "best" models, different link functions were tried and compared:

$$\text{Probit: } \Phi^{-1}(p_i) = \alpha + \beta_i x_{ij}$$
$$\text{C-Log-Log: } log(-log(1-p_i)) = \alpha + \beta_i x_{ij}$$

Comparing the AIC values for both models using all three links:

Logistic	Probit	C-Log-Log
664.7503	664.6108	666.0074
675.2779	674.5091	674.0753

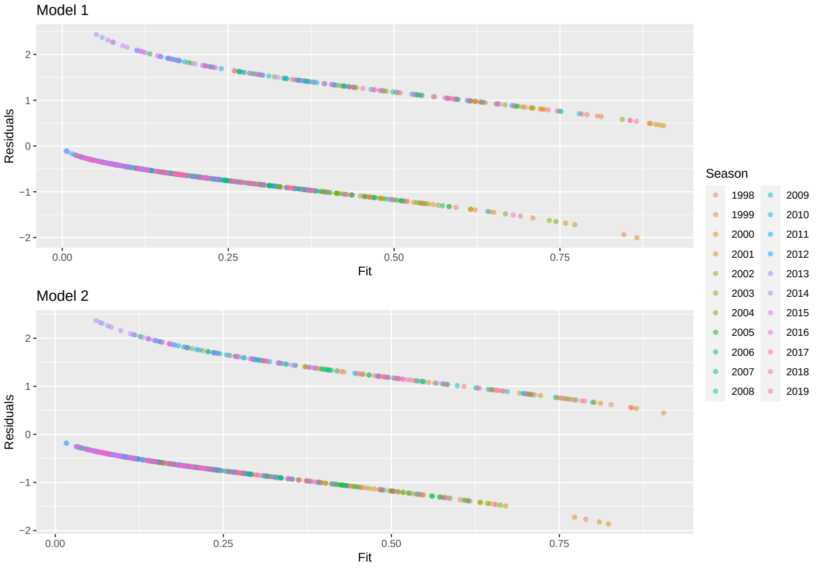
Notes on Model Assumptions

All offensive data for a team in a given season were assumed independent.

Also assumed that each teams performance was independent of other teams performance and a given teams performance season to season was also independent.

The response variable of `Playoff` also has dependent observations.

Model Checking - Residuals



Applying Models to Data

The probability of a team making the playoffs, using both models, is:

$$(1) \qquad p_i = \frac{e^{-115.292+0.1980(R_i)-0.01837(AB_i)-0.00003(R_i * AB_i)}}{1+e^{-115.292+0.1980(R_i)-0.01837(AB_i)-0.00003(R_i * AB_i)}}$$

$$(2) \qquad p_i = \frac{e^{-7.808+87.704(RperAB_i)-19.536(HperAB_i)}}{1+e^{-7.808+87.704(RperAB_i)-19.536(HperAB_i)}}$$

Using these formulas on the 2019 season:

	ARI	ATL	BAL	BOS	CHW
Model 1 Probability	0.3469	0.6563	0.1697	0.2806	0.1706
Model 2 Probability	0.4823	0.6565	0.2320	0.6517	0.1576

Chance of Making Playoffs

For a team with an average team batting average(0.2603):

- A Run Average of .147 achieves 50% chance of going to playoffs
- A Run Average of .172 achieves 90% chance of going to playoffs

For a team with an average team Run Average(0.1345):

- A Batting Average of .204 achieves a 50% chance of going to playoffs
- A Batting Average of 0.3166 achieve a 90% chance of going to playoffs

For a team with an average number of At Bats(5542.671):

- Scoring 806.4 Runs achieves a 50% chance of going to playoffs
- Scoring 937.7 Runs achieves a 90% chance of going to playoffs

Home Run Rates

Using Poisson regression to see which league, which division, and which team has the highest homerun rates.

All models had evidence of over dispersion - with an over dispersion estimate around 7.5 for all models - and AIC values in the 9000's.

The variable of Ball Park Factor was used an offset in all models.

American vs National League

Using a quasipoisson regression:

$$log(\frac{\lambda_i}{BPF_i}) = 0.58768(\mathbf{I}_{AL_i}) + 0.51602(\mathbf{I}_{NL_i})$$

Using observed data it was found:

American League	National League
1.799816	1.675343

Division Comparisons(West vs Central vs East)

Using a quasipoisson regression:

$$
\begin{aligned}
log(\frac{\lambda_i}{BPF_i}) = 0.50845(\mathbf{I}_{AC_i}) &+ 0.64126(\mathbf{I}_{AE_i}) + 0.61208(\mathbf{I}_{AW_i}) \\
&+ 0.55202(\mathbf{I}_{NC_i}) + 0.50342(\mathbf{I}_{NE_i}) \\
&+ 0.48658(\mathbf{I}_{NW_i})
\end{aligned}
$$

Using the observed data, it was found:

	Rates
AL Central	1.6627
AL East	1.8989
AL West	1.8443
NL Central	1.7368
NL East	1.6544
NL West	1.6267

Team Comparisons

Using a quasipoisson model, the homerun rates for each team were modeled.

Using the observed data it was found the top 5 teams were:

NYY	TOR	TEX	CHW	BAL
2.125224	1.953083	1.924496	1.890085	1.888022

Conclusions

The likelihood of a team making the playoffs can be modeled using Offensive and mutated Offensive statistics.

For a team with an average batting average, they would need to score 1.72 runs per ever 10 at bats to have a 90% chance of making the playoffs.

The America League has a higher homerun rate than the national league, with the American League East having the highest rate of all divisions.

An American League East team - New York Yankees - have the highest homerun rate, with 3 of the 5 top teams coming from the American League East and all 5 top teams coming from the American League.