Generalized Linear Models: Induction into the Major League Baseball Hall of Fame

James Henegan

University of Mississippi Medical Center
Department of Biostatistics and Data Science
BDS 726 - Generalized Linear Models

March 4, 2020

Overview

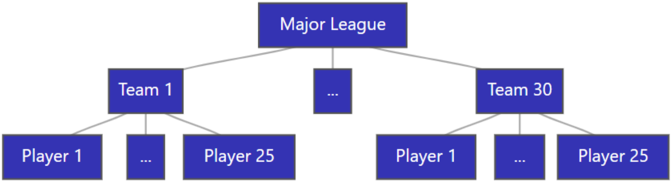
Introduction to Baseball

What is Baseball? Main Ideas and Terminology

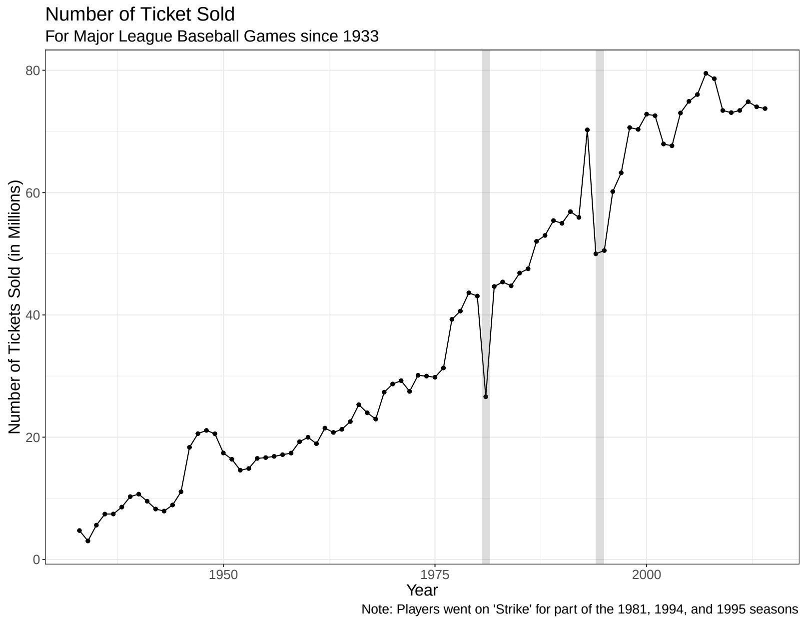
Baseball is a team game where there is offense, defense, and coaching.

- *Batting* is associated with *offense*.
- *Pitching* is associated with *defense*.
- *Managing* is associated with *coaching*.

Professional Baseball Players aspire to play in the *Major League*.



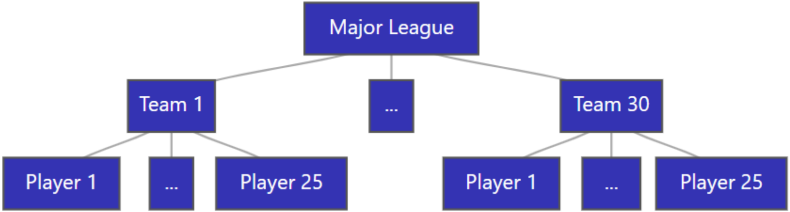
Hierarchical Structure: Players $\subset$ Teams $\subset$ Major League



Number of Ticket Sold
For Major League Baseball Games since 1933
Number of Tickets Sold (in Millions)
0
20
40
60
80
1950
1975
2000
Year
Note: Players went on 'Strike' for part of the 1981, 1994, and 1995 seasons

What is the World Series?

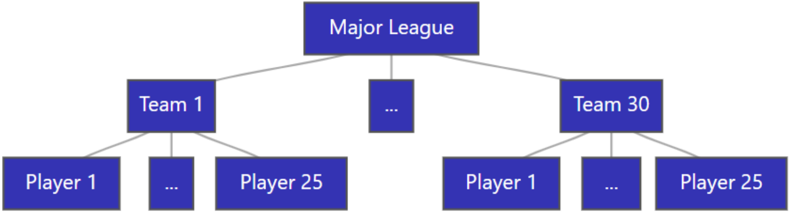
Ever since 1903, at the end of each season, the two best teams in the Major League play against each other in the *World Series.*



Hypothesis: if a manager consistently takes his team to the World Series, then he is probably a pretty good manager.

What is the All-Star Game?

Ever since 1933, the best players from the Major League form two teams and play against each other in the *All-Star Game*.



Hypothesis: if a player consistently appears in the All-Star Game, then he is probably a pretty good player.

What is the Baseball Hall of Fame?

The Baseball Hall of Fame is an organization that commemorates "legendary" players and managers.

Essentially, if a player or manager has an extremely good career, then they may be voted into the hall of fame.

In order to be considered for the Hall of Fame, a player or manager must satisfy the following conditions:

- They must have competed in at least ten seasons.
- They must have retired at least five seasons ago.
- They may not be "banned from baseball" (for cheating, for example).

Research Questions

- How is a manager's "number of World Series appearances" associated with his chances of being inducted into the hall of fame?
- How is a player's "number of All Star Game appearances" associated with his chances of being inducted into the hall of fame?

Outcome Variable

Our outcome variable is "Induction into the Hall of Fame." We will call the outcome variable y. To be explicit, a player or manager will have an y value of `Yes` if he has been inducted into the Hall of Fame, and `No` otherwise.

Grouped Data. When working with grouped data, we will will tally the number of `Yes` values for a group and store the result as `ObsYes` (Observed 'Yes' count). Likewise, we will tally the number of `No` values for a group and store the result as `ObsNo`.

Explanatory Variables

Explanatory Variable for Managers. We will count the number of times each manager took his team to the World Series and record the result as "Number of World Series Appearances" (`numWSA`).

We will group managers by `numWSA`.

Explanatory Variables for Players. We will count the number of times a player was selected to play in the All-Star Game over the course of his career and store the result as `numASG`. We will also record whether the player was a `pitcher` or not (binary variable: 1 for pitcher, 0 otherwise).

We will group players by `numASG` and `pitcher`.

Inclusion and Exclusion Criteria, part 1

Here are some questions we need to consider while forming our data sets:

- Who competed for at least ten seasons?
- Who has been retired for at least five years?
- Who has been banned from baseball?

Inclusion and Exclusion Criteria, part 2

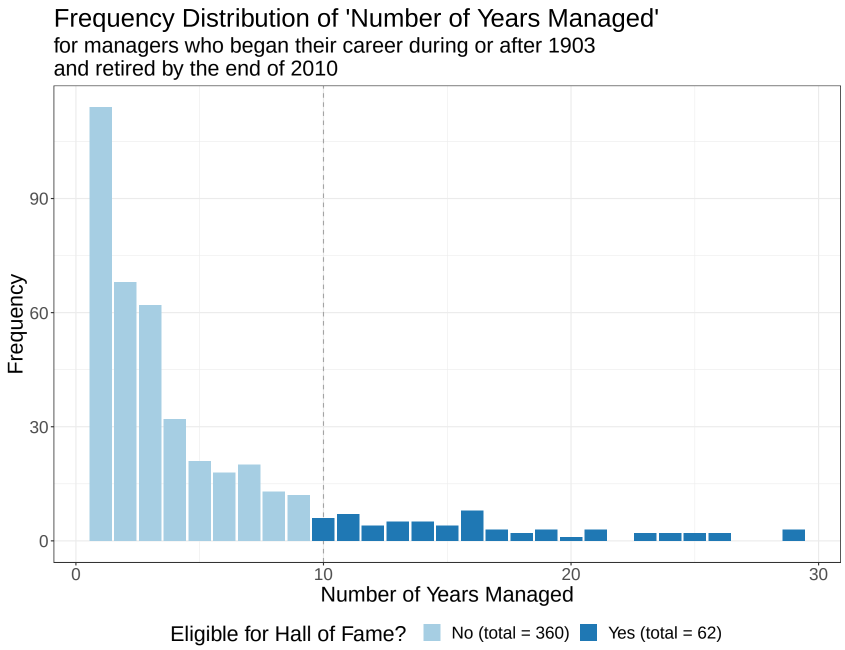
Here are some other issues to consider:

- Should we consider managers who competed before the World Series began in 1903? *Decision: No.*
- Should we consider players who competed before the All Star game started in 1933? *Decision: No.*
- Our primary data source will be Sean Lahman's "History of Baseball" data set, as hosted on Kaggle. It includes Hall of Fame information through the 2015-2016 voting season. Consequently, we will not look at players or managers who competed after the 2010 season.

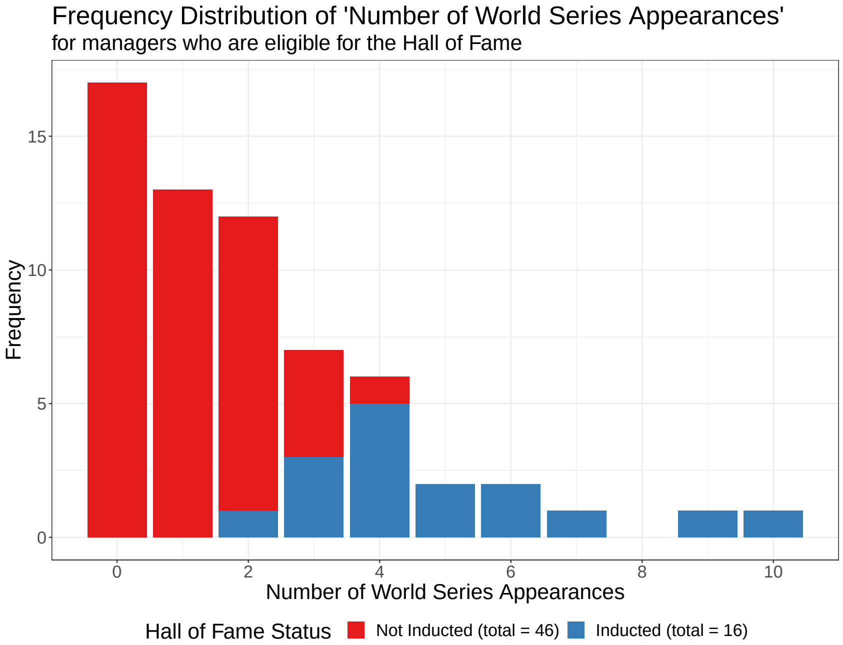
Summary of Introduction

- Batting (offense), pitching (defense), managing (coaching)
- Brief History of Baseball
- World Series, All-Star Game, Hall of Fame
- Research Questions
  - Managers
  - Players
- Outcome Variable: Induction into the Hall of Fame
- Explanatory Variables
  - Managers: World Series Appearances
  - Players: All Star Game Appearances and Position
- Inclusion and Exclusion Criteria

Managers



Frequency Distribution of 'Number of Years Managed'
for managers who began their career during or after 1903
and retired by the end of 2010
Frequency
0
30
60
90
0
10
20
30
Number of Years Managed
Eligible for Hall of Fame?
No (total = 360)
Yes (total = 62)


Frequency Distribution of 'Number of World Series Appearances'
for managers who are eligible for the Hall of Fame
Frequency
0
5
10
15
0
2
4
6
8
10
Number of World Series Appearances
Hall of Fame Status
Not Inducted (total = 46)
Inducted (total = 16)

Observed Data

Number of World Series Appearances	Number of Managers Inducted into the Hall of Fame	Number of Managers Not Inducted into the Hall of Fame	Total
`numWSA`	`obsYes`	`obsNo`	`total`
0	0	17	17
1	0	13	13
2	1	11	12
3	3	4	7
4	5	1	6
5	2	0	2
6	2	0	2
7	1	0	1
9	1	0	1
10	1	0	1
Column Sums	16	46	62

Getting Predictions from the Null Model

$$\text{Let} \quad \hat{\pi} = \frac{\sum \texttt{obsYes}}{\sum \texttt{total}} = \frac{16}{62} \approx 0.258$$

$$\text{Let} \quad \texttt{expYes} = \hat{\pi} \times \texttt{total} \quad \text{and} \quad \texttt{expNo} = (1 - \hat{\pi}) \times \texttt{total}$$

numWSA	obsYes	obsNo	total	$\hat{\pi}$	expYes	expNo
0	0	17	17	0.258	4.39	12.6
1	0	13	13	0.258	3.35	9.65
2	1	11	12	0.258	3.1	8.9
3	3	4	7	0.258	1.81	5.19
4	5	1	6	0.258	1.55	4.45
5	2	0	2	0.258	0.516	1.48
6	2	0	2	0.258	0.516	1.48
7	1	0	1	0.258	0.258	0.742
9	1	0	1	0.258	0.258	0.742
10	1	0	1	0.258	0.258	0.742
	16	46	62			

Evaluating Predictions from the Null Model

For each row, let[1]

$$\texttt{term1} = (\texttt{obsYes}) \times \log\left(\frac{\texttt{obsYes}}{\texttt{expYes}}\right) \quad \text{and} \quad \texttt{term2} = (\texttt{obsNo}) \times \log\left(\frac{\texttt{obsNo}}{\texttt{expNo}}\right)$$

numWSA	obsYes	obsNo	expYes	expNo	term1	term2
0	0	17	4.39	12.6	0	5.07
1	0	13	3.35	9.65	0	3.88
2	1	11	3.1	8.9	-1.13	2.33
3	3	4	1.81	5.19	1.52	-1.04
4	5	1	1.55	4.45	5.86	-1.49
5	2	0	0.516	1.48	2.71	0
6	2	0	0.516	1.48	2.71	0
7	1	0	0.258	0.742	1.35	0
9	1	0	0.258	0.742	1.35	0
10	1	0	0.258	0.742	1.35	0

[1] Note: here, we use the convention that $(0 \times \log 0) = \lim_{x \to 0^+} x \log x = 0$

Evaluating Predictions from the Null Model

numWSA	obsYes	obsNo	expYes	expNo	term1	term2
0	0	17	4.39	12.6	0	5.07
1	0	13	3.35	9.65	0	3.88
2	1	11	3.1	8.9	-1.13	2.33
3	3	4	1.81	5.19	1.52	-1.04
4	5	1	1.55	4.45	5.86	-1.49
5	2	0	0.516	1.48	2.71	0
6	2	0	0.516	1.48	2.71	0
7	1	0	0.258	0.742	1.35	0
9	1	0	0.258	0.742	1.35	0
10	1	0	0.258	0.742	1.35	0

Finally, compute $G^2_{\text{null}} = 2 * \sum(\texttt{term1} + \texttt{term2}) \approx 48.96$.

The quantity G^2_{null} is referred to as the *null deviance*. We will see it again.

Another way to think about the Null Model

In the “null" model, we let $\hat{\pi} = \dfrac{\sum \texttt{obsYes}}{\sum \texttt{total}} \approx 0.258$.

We could have gotten the same estimate $\hat{\pi}$ by "fitting" the model

$$\log \frac{\pi}{1-\pi} = \alpha$$

If we had done it this way, we would have found $\hat{\alpha} = -1.056$.

Then we could write

$$\hat{\pi} = \frac{\exp \hat{\alpha}}{1 + \exp \hat{\alpha}} \approx 0.258$$

Fitting a Conditional Model

Now we consider a "conditional" model

$$\log \frac{\pi}{1-\pi} = \alpha + \beta \times \texttt{numWSA}$$

Fitting this conditional model gives $\hat{\alpha} \approx -7.006$ and $\hat{\beta} \approx 2.215$.

```
---
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -7.006      2.042  -3.430 0.000604 ***
wsa            2.215      0.674   3.287 0.001013 **
---
    Null deviance: 48.95530  on 9  degrees of freedom
Residual deviance:  0.40553  on 8  degrees of freedom
```

A First Look at the Results

Use the residual deviance to evaluate the overall fit of the model.

```
> # Check overall model fit
> # Null hypothesis: our model fits well
> # (compared to the saturated model)
> pchisq(0.40553, 8, lower.tail = F)
[1] 0.9999401
```

Interpreting the Results

Let's interpret the β coefficient. A one unit increase in "number of world series appearances" corresponds to a 2.15 (95% CI: 0.89, 3.54) unit increase in the "log-odds" of being inducted into the hall of fame.

Since it may be difficult to think in terms of the log-odds scale, we can also think about things in terms of odds ratios.

The odds ratio associated with β is $\exp \hat{\beta} \approx 9.16$ (95% CI: 2.45, 34.34).

A one-unit increase in the number of world series appearances corresponds to an 816% increase in the odds of being inducted into the hall of fame (95% CI: 145%, 3334%).

Again, since it may be difficult to think in terms of odds rations, we can also think about things in terms of probabilities.

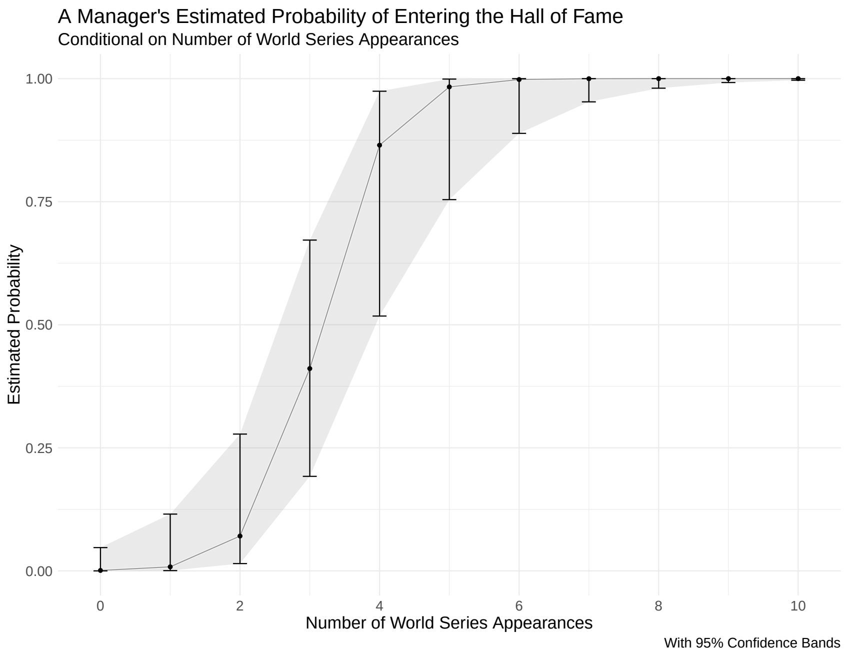
Getting Predictions from the Conditional Model

For each row of the original table, let

$$\hat{\pi} = \frac{\exp(\hat{\alpha} + \hat{\beta} \times \texttt{numWSA})}{1 + \exp(\hat{\alpha} + \hat{\beta} \times \texttt{numWSA})}$$

Let $\texttt{expYes} = \hat{\pi} \times \texttt{total}$ and $\texttt{expNo} = (1 - \hat{\pi}) \times \texttt{total}$

numWSA	obsYes	obsNo	total	$\hat{\pi}$	expYes	expNo
0	0	17	17	0.000906	0.0154	17
1	0	13	13	0.00824	0.107	12.9
2	1	11	12	0.0708	0.849	11.2
3	3	4	7	0.411	2.88	4.12
4	5	1	6	0.865	5.19	0.811
5	2	0	2	0.983	1.97	0.0336
6	2	0	2	0.998	2	0.00372
7	1	0	1	1	1	0.000203
9	1	0	1	1	1	2.42E-06
10	1	0	1	1	1	2.64E-07


A Manager's Estimated Probability of Entering the Hall of Fame
Conditional on Number of World Series Appearances
Estimated Probability
1.00
0.75
0.50
0.25
0.00
0
2
4
6
8
10
Number of World Series Appearances
With 95% Confidence Bands

Evaluating Predictions from the Conditional Model

$$\texttt{term1} = (\texttt{obsYes}) \times \log\left(\frac{\texttt{obsYes}}{\texttt{expYes}}\right) \quad \text{and} \quad \texttt{term2} = (\texttt{obsNo}) \times \log\left(\frac{\texttt{obsNo}}{\texttt{expNo}}\right)$$

numWSA	obsYes	obsNo	expYes	expNo	term1	term2
0	0	17	0.0154	17	0	0.0154
1	0	13	0.107	12.9	0	0.108
2	1	11	0.849	11.2	0.164	-0.15
3	3	4	2.88	4.12	0.125	-0.121
4	5	1	5.19	0.811	-0.185	0.209
5	2	0	1.97	0.0336	0.0338	0
6	2	0	2	0.00372	0.00372	0
7	1	0	1	0.000203	0.000203	0
9	1	0	1	2.42E-06	2.42E-06	0
10	1	0	1	2.64E-07	2.64E-07	0

$$G^2_{\text{residual}} = 2 * \sum(\texttt{term1} + \texttt{term2}) \approx 0.4055267$$

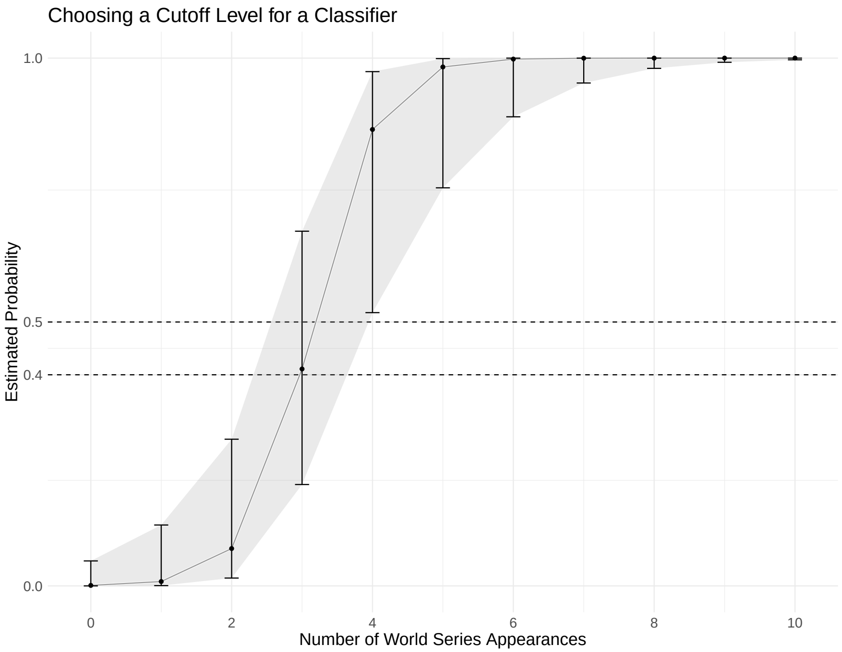
Looking at Deviance

```
    Null deviance: 48.95530  on 9  degrees of freedom
Residual deviance:  0.40553  on 8  degrees of freedom
```

Here's one way to interpret these results: in giving our model knowledge about "World Series appearances" (`numWSA`), we had to spend one degree of freedom. In return, our model was able to bring the deviance down from $G^2_{\text{null}} = 48.96$ to $G^2_{\text{residual}} = 0.41$.

We can test the null hypothesis that the null model fits will compared to the conditional model:

```
> pchisq(q = 48.95530 - 0.40553, df = 1, lower.tail = F)
[1] 3.220121e-12
```


Choosing a Cutoff Level for a Classifier
Estimated Probability
1.0
0.5
0.4
0.0
0
2
4
6
8
10
Number of World Series Appearances

Classifiers and Cutoff Levels

Actual	Prediction, $\pi_0 = 0.5$ $\hat{y} = \texttt{Yes}$	$\hat{y} = \texttt{No}$	Prediction, $\pi_0 = 0.4$ $\hat{y} = \texttt{Yes}$	$\hat{y} = \texttt{No}$
$y = \texttt{Yes}$	12	4	15	1
$y = \texttt{No}$	1	45	5	41

Estimates	$\pi_0 = 0.5$	$\pi_0 = 0.4$
Sensitivity	0.750	0.938
Specificity	0.978	0.891
Accuracy	0.919	0.903

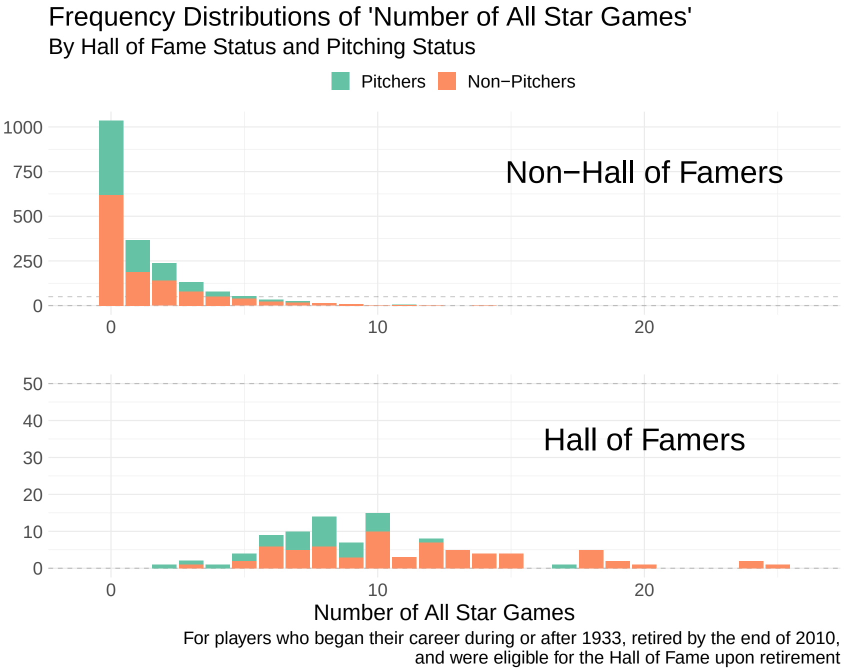
- Sensitivity $= P(\hat{y} = \texttt{Yes} \mid y = \texttt{Yes})$
- Specificity $= P(\hat{y} = \texttt{No} \mid y = \texttt{No})$
- Accuracy $= P(\hat{y} = \texttt{Yes},\ y = \texttt{Yes}) + P(\hat{y} = \texttt{No},\ y = \texttt{No})$

Summary of Manager Model

In this section, we considered a Generalized Linear Model (GLM) which addressed how "Number of World Series Appearances" is associated with a Manager's chances of being inducted into the Hall of Fame

- Structure of GLM
  - Random Component: Binomial
  - Link: Logit Function
  - Systematic Component: Number of World Series Appearances
- Calculated Null Deviance and Residual Deviance "by hand"
- Used the model to estimate probabilities
- Saw how choosing a cutoff level can affect the classification table for a model

Players


Frequency Distributions of 'Number of All Star Games'
By Hall of Fame Status and Pitching Status
Pitchers
Non−Pitchers
Non−Hall of Famers
1000
750
500
250
0
0
10
20
Hall of Famers
50
40
30
20
10
0
0
10
20
Number of All Star Games
For players who began their career during or after 1933, retired by the end of 2010,
and were eligible for the Hall of Fame upon retirement

A Model for the Players

We consider the model

$$\log \frac{\pi}{1-\pi} = \alpha + \beta_1(\texttt{numASG}) + \beta_2(\texttt{pitcher}) + \beta_3(\texttt{numASG} \times \texttt{pitcher})$$

Fitting this model yields the following results:

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.70540    0.54617 -12.277   <2e-16 ***
numASG          0.71925    0.06906  10.415   <2e-16 ***
pitcher        -0.75682    1.01165  -0.748   0.4544
numASG:pitcher  0.37015    0.15756   2.349   0.0188 *
---
    Null deviance: 574.083  on 34  degrees of freedom
Residual deviance:  36.966  on 31  degrees of freedom
```

A Model for the Players

We also consider the model

$$\text{probit}\,\pi = \alpha + \beta_1(\texttt{numASG}) + \beta_2(\texttt{pitcher}) + \beta_3(\texttt{numASG} \times \texttt{pitcher})$$

Fitting this model yields the following results:

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.51263    0.26023 -13.498   <2e-16 ***
numASG          0.37370    0.03435  10.880   <2e-16 ***
pitcher        -0.16959    0.44450  -0.382   0.7028
numASG:pitcher  0.15479    0.07252   2.134   0.0328 *
---
    Null deviance: 574.083  on 34  degrees of freedom
Residual deviance:  33.206  on 31  degrees of freedom

> # overall model fit
> # null hypothesis: model fits well
> pchisq(33.206, 31, lower.tail = F)
[1] 0.3601395
```

Should we simplify the model?

Let's call the model considered on the previous slide the "full" model. Recall that it had a residual deviance of 33.206 with 31 degrees of freedom.

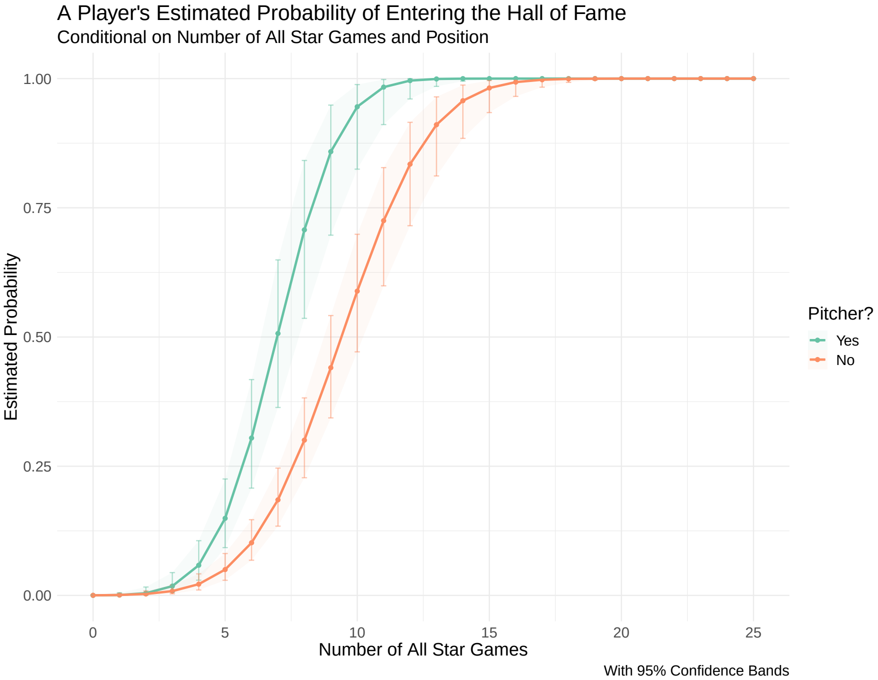
When we fit the "reduced" model

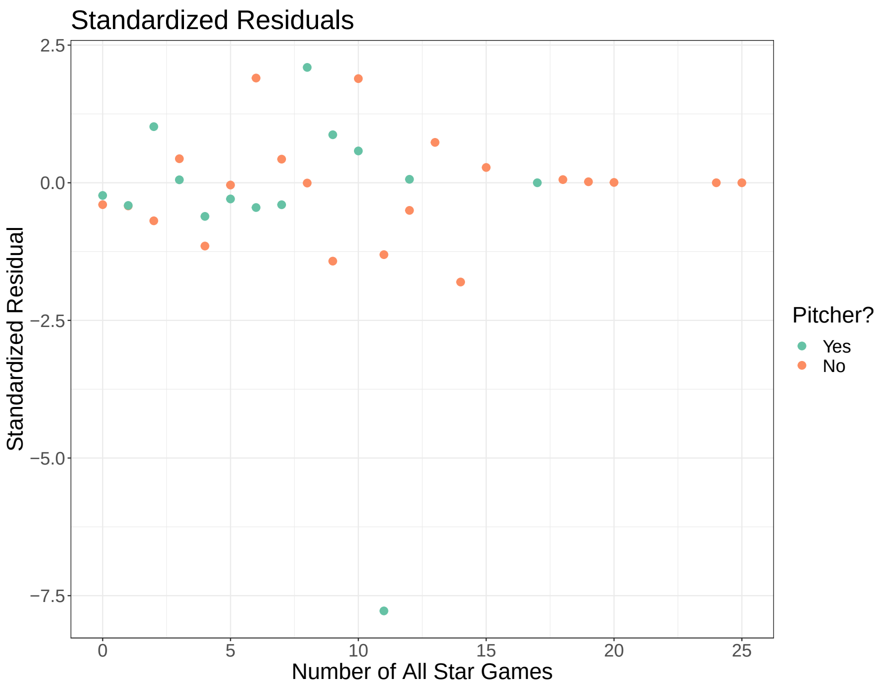
$$\text{probit}\,\pi = \alpha + \beta_1(\texttt{numASG}) + \beta_2(\texttt{pitcher}),$$

we get a residual deviance of 38.517 on 32 degrees of freedom.

This leads to a ΔG^2 of 5.31.

```
> # test the null hypothesis that the reduced model fits well
> pchisq(5.31, 1, lower.tail = F)
[1] 0.02120336
```


A Player's Estimated Probability of Entering the Hall of Fame
Conditional on Number of All Star Games and Position
Estimated Probability
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
25
Number of All Star Games
Pitcher?
Yes
No
With 95% Confidence Bands


Standardized Residuals
Standardized Residual
2.5
0.0
−2.5
−5.0
−7.5
0
5
10
15
20
25
Number of All Star Games
Pitcher?
Yes
No

What caused the largest residual?

Player Name	Position	`numASG`	`inducted`
Jim Bunning	pitcher	9	1
Don Drysdale	pitcher	9	1
Bob Gibson	pitcher	9	1
Rich Gossage	pitcher	9	1
Steve Carlton	pitcher	10	1
Whitey Ford	pitcher	10	1
Tom Glavine	pitcher	10	1
Randy Johnson	pitcher	10	1
Juan Marichal	pitcher	10	1
Roger Clemens[2]	pitcher	11	0
Tom Seaver	pitcher	12	1
Warren Spahn	pitcher	17	1

[2] This player has been accused of using performance enhancing drugs. Could this have affected his chances of being inducted into the Hall of Fame?

Summary of Player's Model

In this section, we considered a Generalized Linear Model (GLM) which addressed how "Number of All Star GameAppearances" and "Position" were associated with a Player's chances of being inducted into the Hall of Fame

- Structure of GLM
  - Random Component: Binomial
  - Link: Probit Function
  - Systematic Component: Number of All Star Game Appearances, Pitcher Indicator
- Performed a Likelihood Ratio Test to see if we could use a simpler model
- Used the model to estimate probabilities
- Examined Standardized Residuals

Conclusion

Conclusion

- Introduction to the game of Baseball
  - World Series
  - All Star Game
  - Hall of Fame
- A GLM on how "Number of World Series Appearances" is associated with a Manager's chances of being inducted into the Hall of Fame
- A GLM on how "Number of All Star Game Appearances" is associated with a Player's chances of being inducted into the Hall of Fame
- The use of grouped data allowed us to look at Deviance Results

Appendix

How to get Standard Errors in the Manager Model?

Let X denote the matrix such that

$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 9 & 10 \end{pmatrix}$$

Note that the bottom row of X^T is our observed values of `numWSA`, while the top row is a “bias" row.

Let W denote the *working weights* of the fit (i.e., the weights in the final iteration of the IWLS fit). These can be obtained in R with `fit$weights`, where `fit` is your fitted model.

Then the variance-covariance matrix between $\hat{\alpha}$ and $\hat{\beta}$ is given by

$$\text{CovMat}(\hat{\alpha}, \hat{\beta}) = (X^T W X)^{-1} = \begin{pmatrix} 4.125393 & -1.3110647 \\ -1.311065 & 0.4492937 \end{pmatrix}$$

Taking the square root of the diagonal terms will give you the standard errors for $\hat{\alpha}$ and $\hat{\beta}$, respectively.

How can we use the Standard Errors to get Confidence Intervals on Predictions?

Let X_0 denote a matrix corresponding to some predictions you'd like to make.

For example, if you'd like to get the standard error for the prediction associated with $\texttt{numWSA} = 4$ and $\texttt{numWSA} = 6$, then you would let $X_0 = \begin{pmatrix} 1 & 4 \\ 1 & 6 \end{pmatrix}$.

Let $C = \text{CovMat}(\hat{\alpha}, \hat{\beta})$. Then the standard error associated with the j-th row of X_0 will be the j-th entry of

$$\text{diag}[(X_0 C X_0^T)^{1/2}].$$

Here, the notation $Y^{1/2}$ means "take the square root of every cell of Y" and $\text{diag}(Y)$ means "extract the diagonal of Y."

Player Stats

Of the 10,005 players who played in the Major Leagues from 1933 to 2010, a total of 2,089 of them played for 10 or more seasons.

Of the 2,089 who played for 10 or more season, 99 were inducted into the baseball hall of fame.

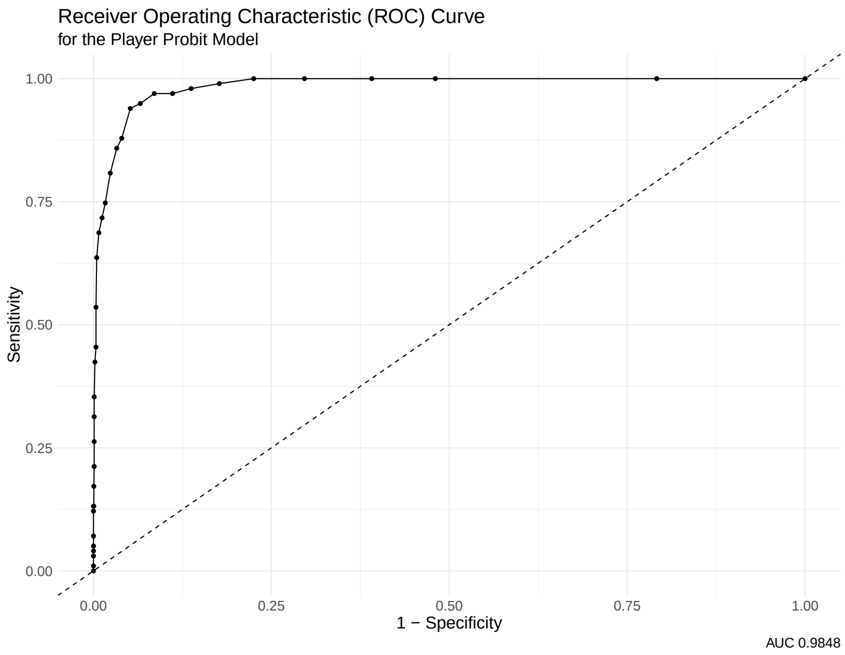
Comparing Logistic Models to Penalized Logistic Models

Managers

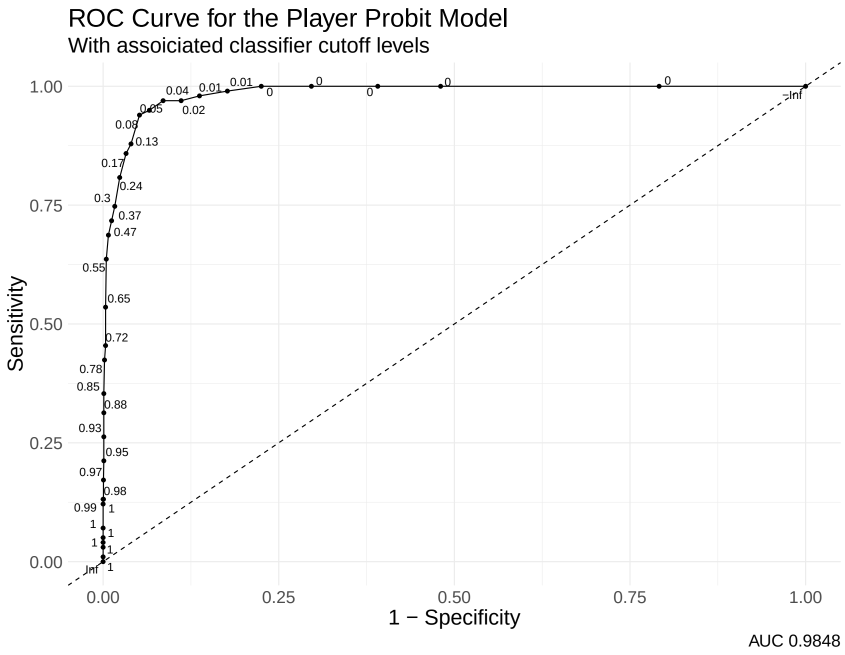
Model	Intercept Coefficient	numWSA Coefficient
Logistic	-7.00569	2.21531
Penalized Logistic	-5.983273	1.878333

Players

Model	Intercept	numASG	pitcher	interaction
Logistic	-6.705399	0.7192467	-0.76	0.37
Penalized Logistic	-6.5916121	0.7057339	-0.62	0.345



Receiver Operating Characteristic (ROC) Curve
for the Player Probit Model
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − Specificity
AUC 0.9848


ROC Curve for the Player Probit Model
With assoiciated classifier cutoff levels
Sensitivity
1 − Specificity
0.00
0.25
0.50
0.75
1.00
AUC 0.9848