Barry Bonds 2001

A look at the greatest single season of any player in baseball history.

By: BK**⊯**

Brett Bejcek and Kyle Voytovich

Brief History

"1961 was the year of Mickey Mantle and Roger Maris. 1998 belonged to Sammy Sosa and Mark McGwire. 2001 saw only one home run king, Barry Bonds."



Source: http://www.baseball-almanac.com/feats/feats0.shtml

During the 2001 Season, what factors played a role in the probability of Barry Bonds getting on base.



- Appearance in the game
- Runners on base
- Inning and outs
- Score at time of at bat
- Opponents ERA



1 season

151 games

648 at bats

11 predictors

Source: http://www.amstat.org/publications/jse/datasets/bonds2001.txt



Exploratory Data Analysis

→ Identify key variables

Based off our intuition of expected relationships.

→ Numerical summaries

Produced pairwise summaries to look at association of variables.

→ Graphical summaries

Translated tables to figures.

→ Model building

Created generalized logistic models based off of previous analysis.



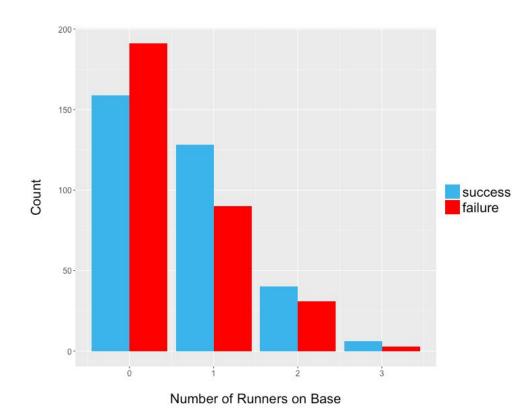


Analysis

During the 2001 season, Barry Bonds showed **no overall difference** in getting on base in home vs. away games.

Runners On Base

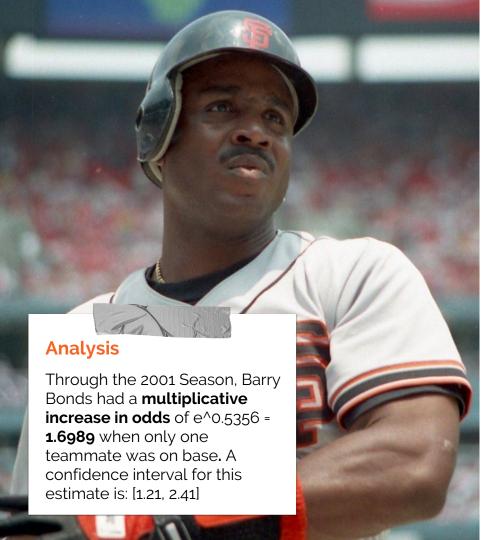
Number on Base	Success	Failure
0	159	191
1	128	90
2	40	31
3	6	3



Effect of Runners On Base

There is a clear leap in Bonds' performance when a runner is on base.

The effect may strengthen for more runners though the data is sparse. Our findings are confirmed in a logistic model.



Model Building

Call:

glm(formula = factor(bonds\$onBase) ~ factor(bonds\$noOnBase),
 family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max -1.482 -1.101 1.032 1.071 1.256

Coefficients:

Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.1834 0.1074 -1.708 0.08762 .
factor(bonds\$no0nBase)1 0.5356 0.1745 3.069 0.00215 **
factor(bonds\$no0nBase)2 0.4383 0.2623 1.671 0.09471 .
factor(bonds\$no0nBase)3 0.8765 0.7152 1.226 0.22037

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 897.82 on 647 degrees of freedom Residual deviance: 886.57 on 644 degrees of freedom

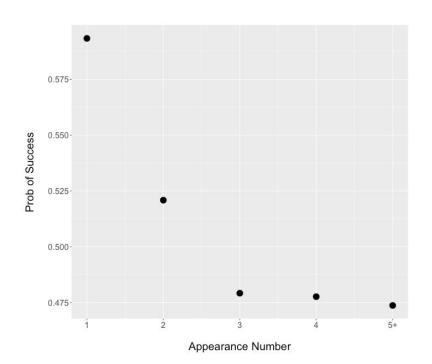
AIC: 894.57

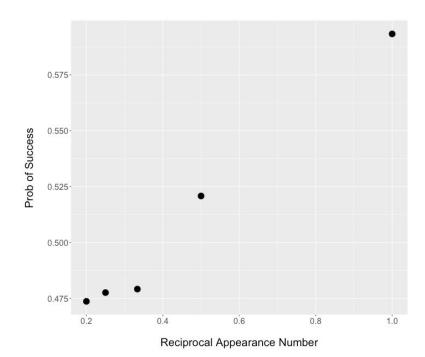
Number of Fisher Scoring iterations: 4

Df Deviance Resid. Df Resid. Dev Pr(>Chi)

NULL 647 897.82 factor(bonds\$no0nBase) 3 11.251 644 886.57 0.01044 *

Appearance in Game





Model Building

```
Call:
glm(formula = factor(bonds$onBase) ~ factor(bonds$anyOnBase) +
    bonds$invApp, family = binomial)
Deviance Residuals:
   Min
            10 Median
                           30
                                  Max
-1.464 -1.104 0.916
                        1.114 1.343
Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)
                         -0.4834
                                     0.1700 -2.843 0.00446 **
factor(bonds$anyOnBase)1
                          0.5159
                                     0.1598 3.228 0.00125 **
bonds$invApp
                          0.6190
                                     0.2708 2.286 0.02227 *
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 897.82 on 647 degrees of freedom
Residual deviance: 881.68 on 645 degrees of freedom
AIC: 887.68
Number of Fisher Scoring iterations: 4
                      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL
                                              897.82
factor(bonds$anyOnBase) 1
                           10.859
                                              886.96 0.0009833 ***
```

5.283

645

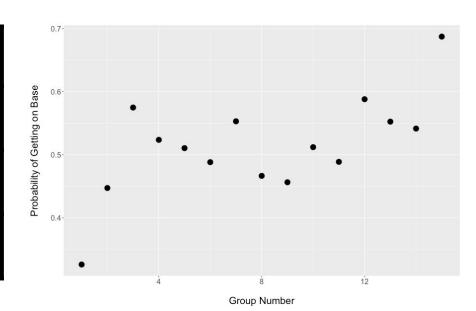
881.68 0.0215348 *

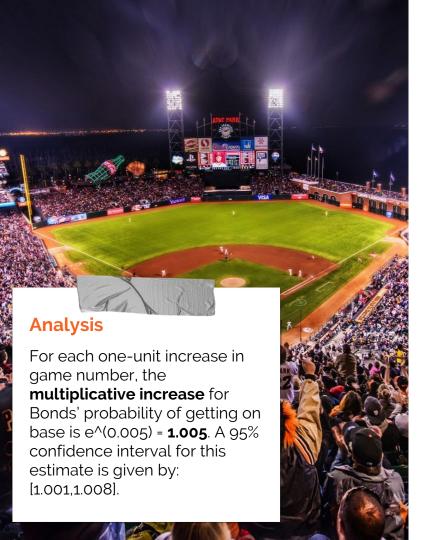
bonds\$invApp



Game (Continuous) Predictor

Subset of Games	Success	Total	% Success
First Third	101	213	47.4%
Middle Third	110	222	49.5%
Last Third	122	213	57.2%





Model Building

Call:

glm(formula = factor(bonds\$onBase) ~ factor(bonds\$anyOnBase) +
bonds\$invApp + bonds\$game, family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max -1.6390 -1.1540 0.7923 1.1433 1.4928

Coefficients:

Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.92003 0.22956 -4.008 6.13e-05 ***
factor(bonds\$anyOnBase)1 0.55526 0.16173 3.433 0.000596 ***
bonds\$invApp 0.63275 0.27295 2.318 0.020440 *
bonds\$game 0.00505 0.00174 2.901 0.003715 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 897.82 on 647 degrees of freedom Residual deviance: 873.14 on 644 degrees of freedom

AIC: 881.14

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			647	897.82		
factor(bonds\$anyOnBase)	1	10.8588	646	886.96	0.0009833	***
bonds\$invApp	1	5.2830	645	881.68	0.0215348	*
bonds\$game	1	8.5398	644	873.14	0.0034747	**



3 Insignificant Variables

Potential predictors that didn't make the cut

- → ERA (Earned Run Average)

 Though opposing pitchers' ERA affected the outcomes of the games, they did little to affect Bonds' batting
- Inning Appearance has a strong effect but inning number does not
- → Number of Outs Bonds seemed to perform better with more outs although the relationship didn't hold in our model (why?)

Let Y_i be whether Bonds gets on base for at bat i in the 2001 season.

Assume

$$Y_i \sim Bern(p_i)$$

where

p_i = prob that Bonds gets on base in at bat i

and

logit
$$p_i = \beta_0 + \beta_1 x_i + \beta_2 y_i + \beta_3 z_i$$

where $x_i = 1$ if someone is on base, 0 otherwise

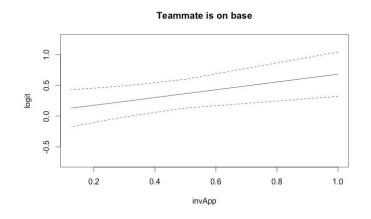
 $y_i = 1 / appearance$

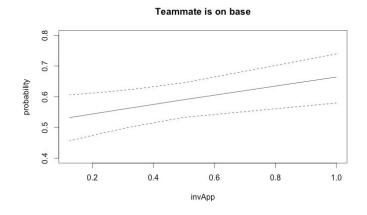
 z_i = game number

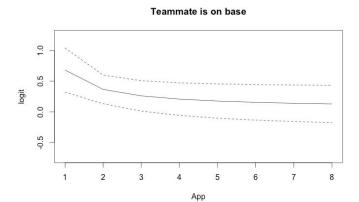
Final Model

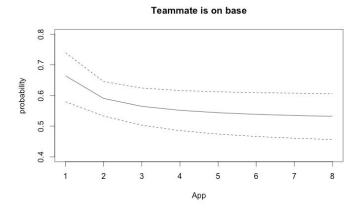
_

Prediction Intervals: Logit vs Prob







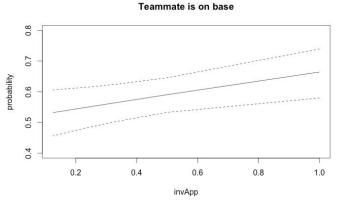


0.8

0.7

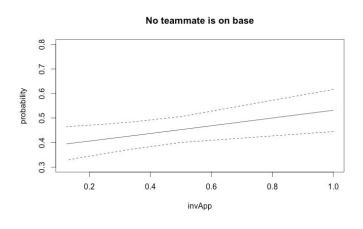
probability

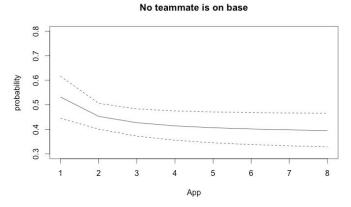
Prediction Intervals: Appearance



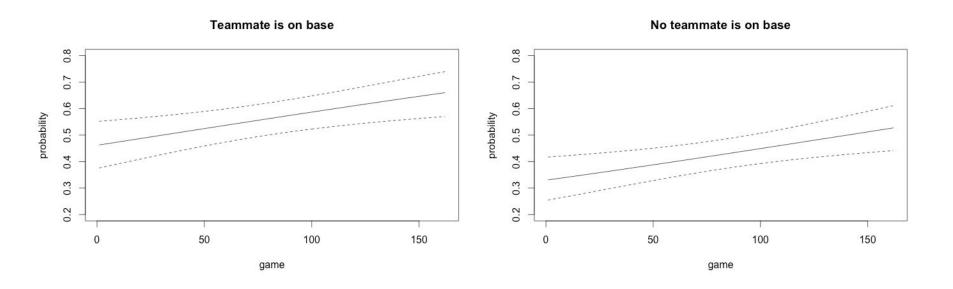
Teammate is on base







Prediction Intervals: Game



Case Predictions

If it is game 30, for Bonds' second appearance and there is no one on base, the probability of Bonds getting on base is:

inv logit(-0.920+(0)(0.555)+(½)(0.633)+(30)(0.005)) =0.3888

If it is game 100, for Bonds' first appearance and there is a teammate on base, the probability of Bonds getting on base is:

inv logit(-0.920+(1)(0.555)+(1)(0.633)+(100)(0.005)) =0.6842



Tip

Use exp(change in logit) to find the multiplicative change in odds.

But inv.logit(logit) to find the probability.

_

Thanks for taking the time to listen to our presentation.

THE END.



Tip

Don't let data stand alone. Always relate it back to a story.