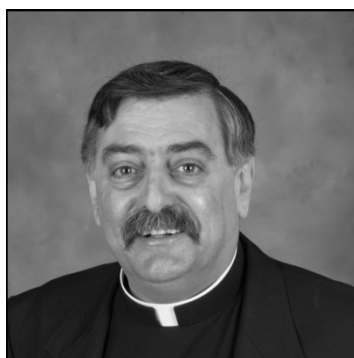


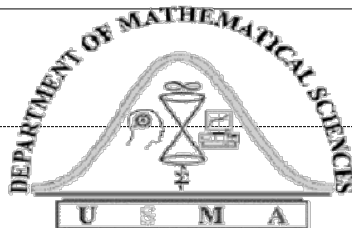
SASabermetrics: PROC LOGISTIC in the Hall of Fame



ROD STURDIVANT
BASUG SEPTEMBER, 2010



Father Gabe Costa, Ph.D
Professor of Mathematics
Department of Mathematical Sciences
West Point



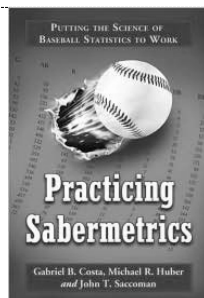
Father Gabe with Babe Ruth's granddaughter



MA488: Sabermetrics

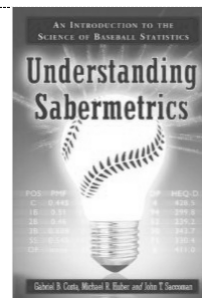
Goals and Objectives: This course is designed to advance your understanding of statistical concepts and to develop your skills in applying those statistics to baseball.

SABERMETRICS

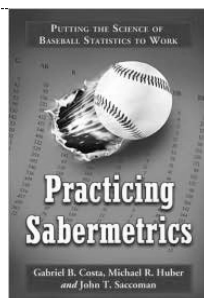


**Who is the
GREATEST
player/
hitter/X of
all time?**

*“The search for
objective
knowledge about
baseball”*

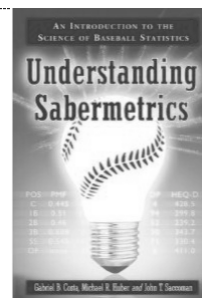


SABERMETRICS



**Who is the
GREATEST
player/
hitter/X of
all time?**

*“The search for
objective
knowledge about
baseball”*



**If your
analysis
doesn't say
Babe Ruth...
it is wrong.
Fr. Costa**

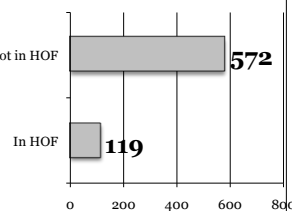
The BASEBALL HALL OF FAME Cooperstown, New York

- Getting in:
 - Nominated (2 of 6 committee or 5% of ballots previous year)
 - BWAA members vote for 0 to 10
 - 75% of ballots required
 - Retired 5-20 years (played in 10)
- 292 elected members.
 - 203 former major league players
 - Baseball Writers' Association of America has elected 109 candidates
 - Veterans Committee 157
 - 16 catchers
 - 21 first basemen
 - 19 second basemen
 - 14 third basemen
 - 23 shortstops
 - 21 left fielders
 - 23 center fielders
 - 24 right fielders



DATA

- OFFENSIVE data only
 - More than 20 measures (statistics)
 - Measures of power (HR, SLG PCT...), speed (SB, triples), consistency (average, strike outs, walks), team play (sacrifice hits)
- Post-1900 players only (last year 2001)
- More than 4,000 career at bats (HOF eligible)
- www.retrosheet.org



watch?v=4ltD21rYVWw



A (quick and dirty) Literature Review

- Several books/articles with sabermetric measures to predict Hall of Fame (Bill James most notable)
- Findlay and Read (2002) closest:
 - Fit logistic and linear regressions
 - Pre-selected predictors
 - Data from 1967-1993 (players with at least 1 vote – 67 players)
 - Small prediction/test sets (11 players – 7 in data)
 - More than just offensive statistics
 - ✦ Defense: position (3 categories), gold gloves
 - ✦ Honors/awards: MVP, ROY etc.
 - ✦ Race
 - ✦ League (AL and NL)



THE BASIC LOGISTIC REGRESSION MODEL

y is the binary response (in HOF = 1, not in HOF = 0), and we have p predictors in a model for the PROBABILITY of making the HOF (π)

$$y = \pi(\underline{x}) + \varepsilon$$

$$E(\varepsilon) = 0$$

$$\text{var}(\varepsilon) = \pi(\underline{x})[1 - \pi(\underline{x})]$$

$$\pi(\underline{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}$$

Solving for the probability yields the linear function of the “log odds” or “logit”

$$g(\underline{x}) = \ln \left\{ \frac{\pi(\underline{x})}{1 - \pi(\underline{x})} \right\} = \beta_0 + \beta_1 x + \dots + \beta_p x$$

AFTER (more than just) A FEW HOURS MODEL BUILDING...

- Preliminary variable selection using STEPWISE and “PURPOSEFUL” approaches
- Issues with correlation in covariates (i.e. $H = 1b + 2b + 3b + HR$) – no impact on final model stability, but lead to multiple models

MODEL 1 (forward)	MODEL 2 (purpose)
Position (POS)	POS
Runs (R)	R
Doubles (2B)	Triples (3B)
Runs Batted In (RBI)	Home Runs (HR)
Batting Average (AVG)	On base PCT (OBP)
Strikeouts (SO)	SO
	Walks (BB)
	Hit by Pitch (HBP)
	Sacrifices (SAC)

← “DEFENSE”
 ← “SPEED”
 ← “POWER”
 ← “AVERAGE”
 ← “TEAM PLAY”

Some SAS Code/Output

```
PROC LOGISTIC DATA = post1900 desc;
  CLASS pos bats / param = ref ;
  MODEL hof = G pos slg nobp navg sh hbp bb sb rbi
             hr _b1 _b0 _b h r ab bats so / selection = f ;
run ;
```

The LOGISTIC Procedure

Summary of Forward Selection

Model Information				Summary of Forward Selection			
				Step	Effect	Number	Score
Data Set	WORK.POST1900			1	R	1	238.5980
Response Variable	HOF			2	navg	1	34.6444
Number of Response Levels	2			3	pos	6	58.7552
Model	binary logit			4	RBI	1	9.2350
Optimization Technique	Fisher's scoring			5	_B0	1	15.4969
				6	SO	1	8.1824
Number of Observations Read	694						
Number of Observations Used	691						

Response Profile

Ordered Value	HOF	Total Frequency
1	1	116
2	0	575

Probability modeled is HOF=1.

MODEL CHOICE?

- Rule of thumb (for PREDICTION): choose most parsimonious model...

CHOICES	AIC	SC	-2 Log Lik	C	R ²	H-L
MODEL 1 – 6 variables POS R SO RBI 2B AVG	303.1	357.6	279.1	0.947	0.402	0.62
MODEL 2 – 9 variables POS R SO HR 3B OBP HBP BB SAC	307.5	375.5	277.5	0.945	0.404	0.19

- So I chose...

MODEL 1

Note: here something like cross-validation could have been useful; note there is not much difference in models by most measures

Some SAS Code/Output

```
PROC LOGISTIC DATA = post1900 desc ;
  CLASS pos bats / param = ref ;
  MODEL hof = r pos so rbi navg _b0 ;
run ;
```

The LOGISTIC Procedure

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	636.863	303.108
SC	641.401	357.565
-2 Log L	634.863	279.108

R-Square 0.4024 Max-rescaled R-Square 0.6696

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	355.7551	11	<.0001
Score	323.5513	11	<.0001
Wald	118.2255	11	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
pos	6	44.6579	<.0001
R	1	49.1015	<.0001
SO	1	7.5851	0.0059
navg	1	20.6482	<.0001
RBI	1	31.9585	<.0001
_B0	1	19.1174	<.0001

All predictors highly
significant

(You can think of many I am sure) Modeling Improvements



- Other measures:
 - DEFENSIVE STATISTICS
 - OTHER SABERMETRIC MEASURES
 - AWARDS (ROY, MVP, Gold Glove etc)
 - “CLUTCH” PLAY (post-season data)
- Other approaches/additional work
 - INTERACTIONS
 - SCALE of COVARIATES
 - FACTOR ANALYSIS FIRST
 - CROSS VALIDATION

...but let's see how we did anyway!

ASSESSING MODEL FIT

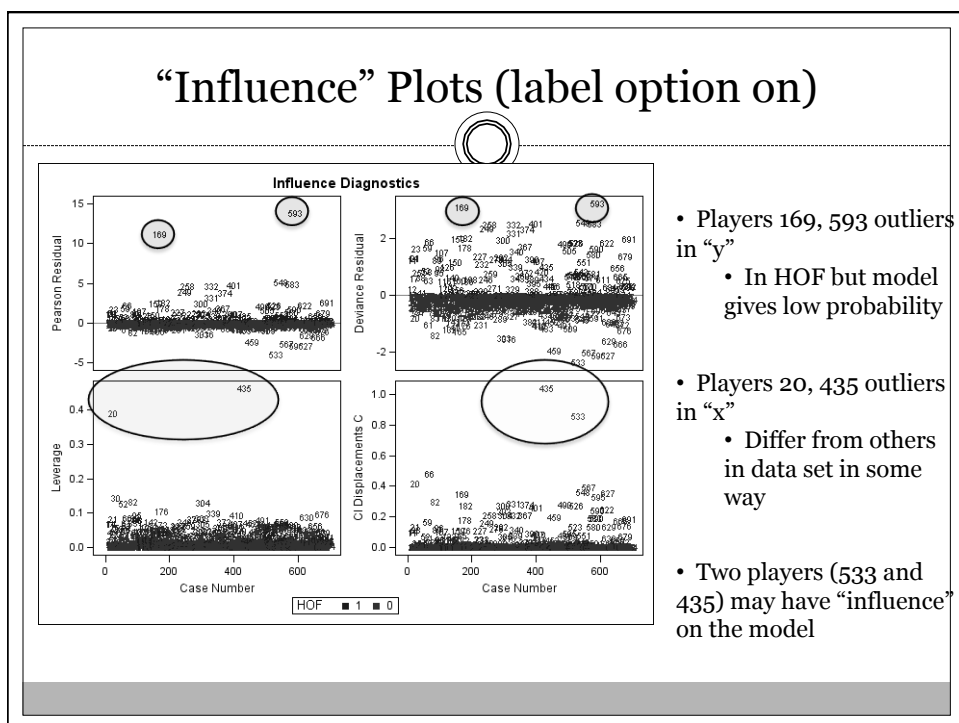
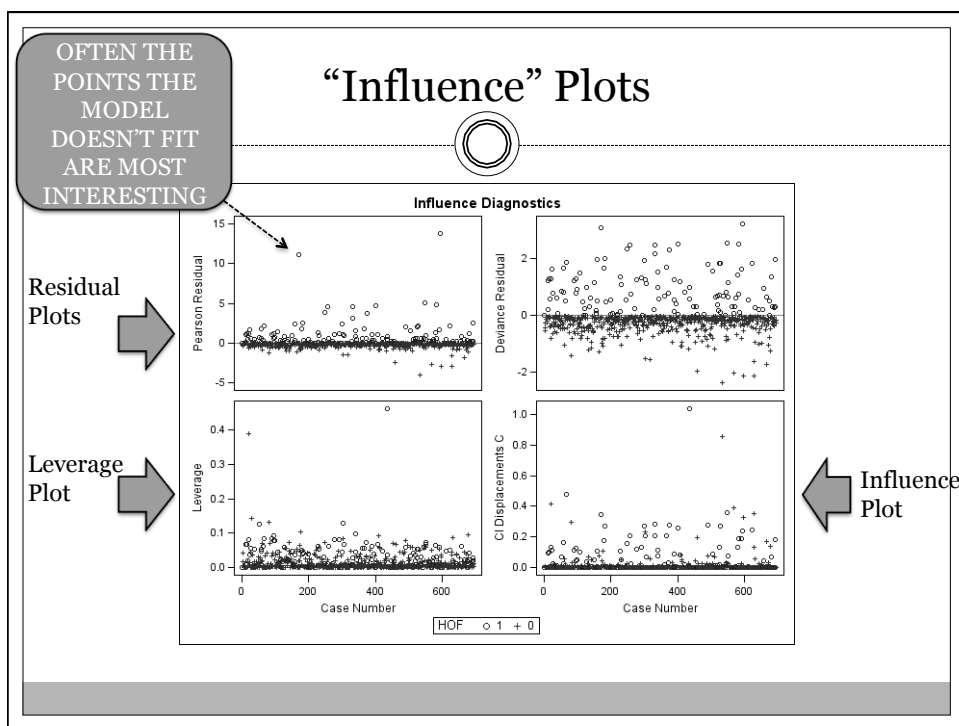


1. Diagnostics (NOT usually where I start)

```
ods graphics on ;
PROC LOGISTIC DATA = post1900 desc
    PLOTS(label)= (phat leverage influence dpc ) ;
    CLASS pos bats / param = ref ;
    MODEL hof = r pos so rbi navg _b0 ;
run ;
ods graphics off ;
```

**SAS MAKES THIS VERY EASY WITH SOME NICE
ods graphics
PLOTS**

“label” option can help identify interesting subjects



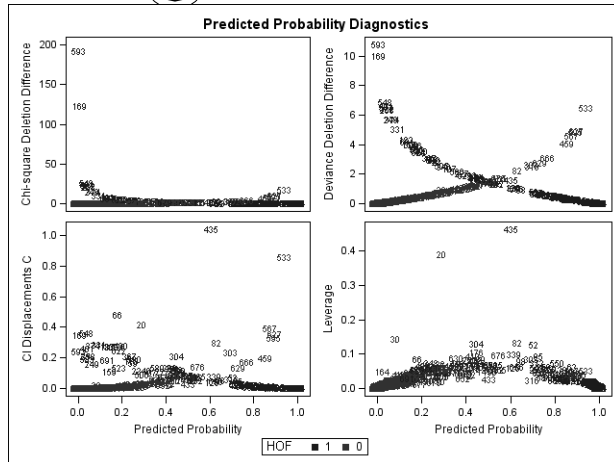
“phat” plots

“Influence” plots (1 leverage) with predicted probability as the x-axis.

Same 5 players as on previous plots

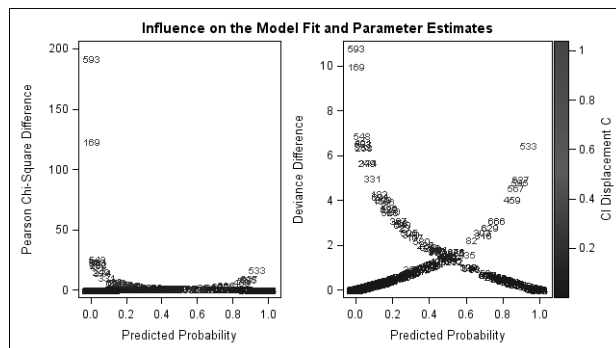
Note:

3 in HOF BUT model predicts shouldn't be



INFLUENCE (from the “dpc” plot)

Same idea – I like the colored bar to give an idea about how influential on the right. In logistic regression, very much a “rule of thumb” though...collect suspicious points and examine them:



5 PLAYERS:

Leverage 20 435

Residual 169 593

Influence 533

Any
GUESSES
?

Influential Players?

Leverage (different “x”):

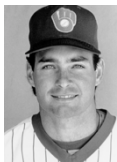


HAROLD BAINES

- 22 year playing career
- Numbers comparable to Jim Rice
- BA 0.289, 384 HRs
- 6 All-star games
- Hall of Fame votes 5-6% 2007-2010

PAUL MOLITOR

- 21 year playing career
- 9th all time in hits (3,319)
- BA 0.306, 234 HRs
- 7 All-star games
- 1993 World Series MVP
- Elected to Hall of Fame 2004 (85%)



High Residual (“y”):

“The Lip” LEO DUROCHER

- “Nice guys finish last.”
- 17 year playing career – 2b/SS
- Good defensively
- BA 0.247, 24 HRs
- 3 All-star games
- Elected to Hall of Fame 1994 (vets)



CASEY STENGEL “Old Professor”

- “All right, everybody line up alphabetically according to your height.”
- 14 year playing career – OF
- BA 0.284, 60 HRs
- Led league in OBP once
- Elected to Hall of Fame 1966 (vets)



High Influence

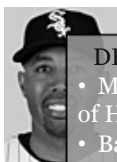


PETE ROSE “Charlie Hustle”

- 24 year playing career – OF/1b/3b
- BA 0.303 (led league 3 times)
- 17 All-star games, ROY
- MVP of league and W.S. 1 time
- 2 Gold Gloves
- FIRST All-time: AB, Games, Hits (4256), Singles
- 2nd all-time doubles, 6th in runs
- Not in Hall of Fame – 9.5% 1992, < 5% in 1993 and 1994

Influential Players?

Leverage (different “x”):



HAROLD BAINES

DESIGNATED HITTERS

- Molitor 0.61 probability of HOF
- Baines 0.28
- 4 other DH in data set all “average” (less than 0.10): (Baylor, Downing, McRae, Thornton)

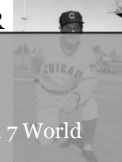
- 21 year playing career
- 9th all time in hits (3,319)
- BA 0.306, 234 HRs
- 7 All-star games
- 1993 World Series MVP
- Elected to Hall of Fame 2004 (85%)



High Residual (“y”):

“The Lip” LEO DUROCHER

- “Nice guys finish last.”
- 17 year playing career – 2b/SS
- Good defensively
- BA 0.247, 24 HRs
- 3 All-star games
- Elected to Hall of Fame 1994 (vets)



CASEY STENGEL “Old Professor”

- “All right, everybody line up alphabetically according to your height.”
- 14 year playing career – OF
- BA 0.284, 60 HRs
- Led league in OBP once
- Elected to Hall of Fame 1966 (vets)



High Influence



PETE ROSE “Charlie Hustle”

- Model: 0.94 probability of making Hall of Fame
- 38th highest ranked player in data set
- 24 year playing career – OF/1b/3b
- BA 0.303 (led league 3 times)
- 17 All-star games, ROY
- MVP of league and W.S. 1 time
- 2 Gold Gloves
- FIRST All-time: AB, Games, Hits (4256), Singles
- 2nd all-time doubles, 6th in runs
- Not in Hall of Fame – 9.5% 1992, < 5% in 1993 and 1994

Impact on Model

- Removing Baines/Molitor only impacts the DH group (lowers probability, but only 4 players left)
- Removing Rose minimal impact on model
- Removing Stengel/Durocher minimal impact on model

...in general, removing improves fit slightly

I removed STENGEL and DUROCHER from data
– MANAGERS not a part of the project

ROSE tempting, but “character” a part of HOF so
left him in (hurts model)

ASSESSING MODEL FIT

2. Summary measures (we saw many during model fitting) – using the “rsquare” and “lackfit” options in the model statement:

The LOGISTIC Procedure				Partition for the Hosmer and Lemeshow Test					
Model Fit Statistics				Group	Total	HOF = 1		HOF = 0	
Criterion	Intercept Only	Intercept and Covariates	Observed			Expected	Observed	Expected	
				1	69	0	0.04	69	68.96
				2	69	0	0.12	69	68.88
				3	69	0	0.26	69	68.74
AIC	629.799	284.553		4	69	0	0.52	69	68.48
SC	634.334	352.581		5	69	0	1.00	69	68.00
-2 Log L	627.799	254.553		6	69	5	2.31	64	66.69
				7	69	3	4.73	66	64.27
R-Square	0.4183	Max-rescaled R-Square	0.6995	8	69	11	11.27	58	57.73
				9	69	34	33.55	35	35.45
Testing Global Null Hypothesis: BETA=0				10	68	64	63.20	4	4.80
Test	Chi-Square	DF	Pr > ChiSq	Hosmer and Lemeshow Goodness-of-Fit Test					
Likelihood Ratio	373.2462	14	<.0001						
Score	335.9620	14	<.0001						
Wald	116.4922	14	<.0001						
				Chi-Square		DF	Pr > ChiSq		
				6.0400		8	0.6428		

CLASSIFICATION

- Using the “ctable” model option, look at 0.5 cut point:

Prob Level	Correct		Incorrect		Correct	Percentages			
	Event	Non-Event	Event	Non-Event		Sensitivity	Specificity	False POS	False NEG
0.500	79	553	19	38	91.7	67.5	96.7	19.4	6.4

OBSERVED			
CLASSIFIED	In HOF	Not in HOF	Total
Make HOF	79	19	98
Don't make HOF	38	553	591
Total	117	572	689

$79/117 = 67.5\%$ SENSITIVITY
 $553/572 = 96.7\%$ SPECIFICITY

Overall: 91.7% correct classification

TRADE-OFF

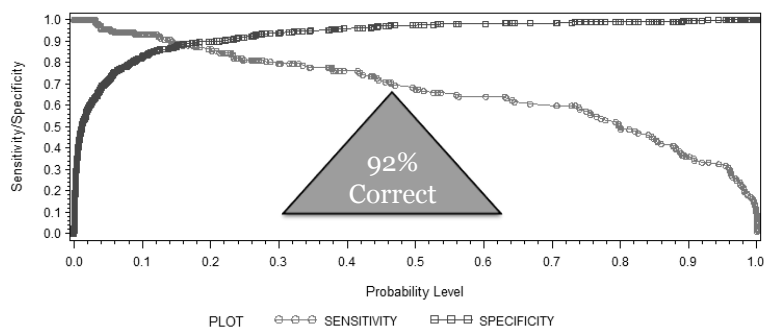
Raising cut-point

Helps SPECIFICITY
(fewer predicted
HOF)

Hurts SENSITIVITY

CLASSIFICATION

SENSITIVITY and SPECIFICITY



Predict
ALL make
HOF

Predict
NOONE
makes HOF

CLASSIFICATION

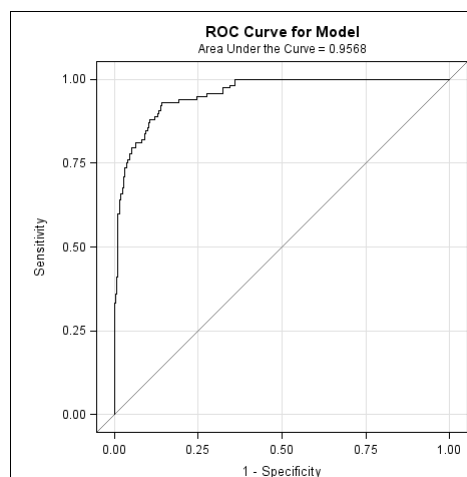
Model has excellent discrimination:

Area under ROC > 0.9

Appears probability cut-point around 0.5 (or slightly less) not bad

-Possibly better if consider POSITION:

- 2b (0.45 slight improvement)
- C (0.4 slight improvement)
- OF (0.4 slight improvement)
- SS...hard to discriminate!
- 3B easiest (gap from 0.46 to 0.75)



EXTERNAL VALIDATION

- Players on ballot 2008, 2009, 2010

In data set

16/19=84%

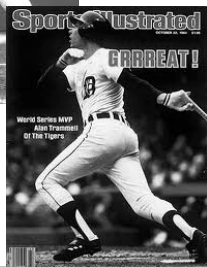
Not in data set

Player	Ballots (high)	Model
Andre Dawson	77.9%	0.512
Jim Rice	76.4%	0.801
Alan Trammell	22.4%	0.729
Mark McGwire	23.7%	0.436
Don Mattingly	16.1%	0.148
Dave Parker	15.2%	0.238
Harold Baines	6.1%	0.271
Dale Murphy	13.8%	0.099
Dave Concepcion	16.2%	0.075



Player	Ballots (high)	Model
Rickey Henderson	94.8%	0.986
Roberto Alomar	73.7%	0.935
Barry Larkin	51.6%	0.850
Edgar Martinez	36.2%	0.076
Tim Lincecum	30.4%	0.719
Fred McGriff	21.5%	0.463
Andres Galarraga	4.1%	0.102
Robin Ventura	1.3%	0.072
Ellis Burks	0.4%	0.328
Mark Grace	4.1%	0.152

You be the Judge



Model struggles to
predict
SHORTSTOPS
They may still get in!

- **Barry Larkin (SS)**

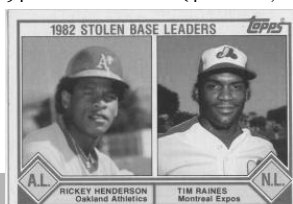
Stats comparable to HOFers Reese, Cronin, Sandberg, (Jeter) & fares well with most HOF metrics
12 time All-Star, 3 Gold Gloves, 1 MVP
19 seasons, 0.295 hitter with 198 HRs at shortstop

- **Alan Trammell (SS)**

Stats comparable to Larkin
6 time All-Star, 4 Gold Gloves, 1 World Series MVP
20 seasons, 0.285 hitter with 185 HRs at shortstop

- **Tim Lincecum (OF)**

Stats comparable to HOFers Brock, Slaughter and metrics close to average HOF member
7 time All-Star (1 MVP), led league in AVG (1), OBP (1), R (2)
23 seasons, 0.294 hitter with 808 SBs (4th all time, led league 4 times)



Model Based Inference

- **ODDS of MAKING HALL of FAME**



FRED LYNN
5.5% (1996)

9 time all-star, 306 HR,
4 Gold Gloves, ROY
• MODEL: $P(\text{HOF}) = 0.065$

• AVERAGED 90 RUNS
first 5 years as Red Sox
only 57 in next 11 years

SUPPOSE HE SCORED
300 MORE RUNS IN
THOSE YEARS?

Odds Ratio (and 95%
confidence interval)
for 300+ R
5.86 (3.17, 10.85)

3 time all-star
442 HR 35th All-time
16 seasons, 162 game
averages of 37 HR and 101
RBI

• MODEL: $P(\text{HOF}) = 0.006$

Hit 0.235 and struck out
1816 (11th all time) times!

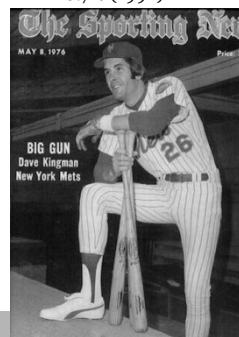
Odds Ratio for 10 point
higher BA:

1.76 (1.39, 2.22)

Odds Ratio for 100 fewer
strikeouts:

1.19 (1.05, 1.35)

DAVE KINGMAN
0.7% (1992)



MORE ODDS



• Who is the Hall of Famer?

Player 1

Player 2

- | | |
|---------------------|---------------------|
| • 20 seasons | • 24 seasons |
| • 0.272 | • 0.269 |
| • 385 HR | • 376 HR |
| • 1384 RBI | • 1330 RBI |
| • 1470 Runs | • 1276 Runs |
| • 0.470 SLG pct | • 0.457 SLG pct |
| • 0.370 OBP | • 0.341 OBP |
| • 3 time All-star | • 11 time All-star |
| • 8 Gold Gloves | • 1 Gold Gloves |
| • 2 Silver Sluggers | • 3 Silver Sluggers |
| | • ROY |

MORE ODDS



• Who is the Hall of Famer?

Player 1

Player 2

- | | |
|---------------------|---------------------|
| • 20 seasons | • 24 seasons |
| • 0.272 | • 0.269 |
| • 385 HR | • 376 HR |
| • 1384 RBI | • 1330 RBI |
| • 1470 Runs | • 1276 Runs |
| • 0.470 SLG pct | • 0.457 SLG pct |
| • 0.370 OBP | • 0.341 OBP |
| • 3 time All-star | • 11 time All-star |
| • 8 Gold Gloves | • 1 Gold Gloves |
| • 2 Silver Sluggers | • 3 Silver Sluggers |



1 - DWIGHT EVANS
10.4%
Model 0.303

2 - CARLTON FISK
Hall of Fame (2000, 76%)
Model 0.908



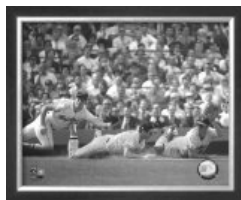
Odds ratio for C vs. OF:
40.3 (10.2, 158.6)
Evans probability if C: 0.827

Model or Voters?

In HOF

Probability < 0.2

Schalk	Ray	C	0.032
Southworth	Billy	OF	0.032
Mazeroski	Bill	2B	0.035
Harris	Bucky	2B	0.039
Kelly	George	1B	0.039
Hafey	Chick	OF	0.053
Lopez	Al	C	0.059
Kell	George	3B	0.088
Youngs	Ross	OF	0.122
Doby	Larry	OF	0.126
Ferrell	Rick	C	0.126
Tinker	Joe	SS	0.142
Huggins	Miller	2B	0.144
Rizzuto	Phil	SS	0.152
Robinson	Brooks	3B	0.179
Brock	Lou	OF	0.185



Brooks Robinson



Evidence of model missing
DEFENSE
(Robinson, 2b/SS) and ERA

NOT In HOF

Probability > 0.5

Minoso	Minnie	OF	0.500
Cramer	Doc	OF	0.527
Wills	Maury	SS	0.527
Randolph	Willie	2B	0.533
Doyle	Larry	2B	0.564
Bush	Donie	SS	0.603
Johnson	Bob	OF	0.707
Jackson	Joe	OF	0.707
Trammell	Alan	SS	0.729
Whitaker	Lou	2B	0.801
Myer	Buddy	2B	0.877
Stephens	Vern	SS	0.890
Simmons	Ted	C	0.897
Torre	Joe	C	0.920
Rose	Pete	OF	0.959

"CHARACTER" criteria –
"Shoeless" Joe Jackson
Pete Rose

Model v. Votes

Hall of Famers

Bill Mazeroski, 2B (p = 0.035)
 • 17 years
 • 8 Gold Gloves, 10 All-star, 1 player of year
 • 0.260 AVG (best 0.283), 0.299 OBP
 • 138 HR (best 19), 2016 H
 • Best RBI 82, R 71
 HOF Votes: 6.1% ('78)-42% ('92)
 2001 Vet pick



Tony Perez, 1B (p = 0.2)
 • 23 years
 • 7 All-star (1 MVP)
 • 0.279 AVG (best 0.344 OBP, 0.463 SLG)
 • 379 HR (best 40, 2 years 30+), 2732 H
 • 1,652 RBI (7 years 100+)
 HOF Votes: 50% ('92)-77.2% ('00);



Model Prefers

Maury Wills, SS (p = 0.52)
 • 14 years
 • 2 GG, 7 AS (1 MVP), 1 POY, 1 NL MVP
 • 0.281 AVG, 0.330 OBP
 • 586 SB (19th), led league 6 times (best 104)
 • 2134 H, 1067 R (best 130)
 HOF Votes: 30.3% ('78)-25.6% ('92); highest vote 40%



Gil Hodges, 1B (p = 0.3)
 • 18 years player
 • 3 Gold Gloves, 8 All-star
 • 0.273 AVG (best 0.304), 0.359 OBP, 0.487 SLG
 • 370 HR (best 42, 6 30+), 2016 H
 • 1,274 (7 years 100+) RBI
 HOF Votes: 24% ('69)-63% ('88);
 MIRACLE METS (9 years as a manager)



The Real (Fr. Costa) model validation test



Hall of Fame
All Time Highest Vote
Percentages



MODEL
Highest Probabilities

NAME	POS	YEAR	% VOTES	RANK
Cal Ripken, Jr.	SS	2007	98.53%	3
Ty Cobb	CF	1936	98.23%	4
George Brett	3B	1999	98.19%	5
Hank Aaron	RF	1982	97.83%	6
Tony Gwynn	RF	2007	97.61%	7
Mike Schmidt	3B	1995	96.52%	8
Johnny Bench	C	1989	96.42%	9
Babe Ruth	RF	1936	95.13%	11
Honus Wagner	SS	1936	95.13%	
Rickey Henderson	LF	2009	94.81%	13
Willie Mays	CF	1979	94.68%	14
Carl Yastrzemski	LF	1989	94.63%	15
Reggie Jackson	RF	1993	93.62%	17
Ted Williams	LF	1966	93.38%	18

Model: Jackson 0.46 (others > 0.84)

NAME	POS	Probability	HOF %
✓ Babe Ruth	RF	99.99%	
Ty Cobb	CF	99.99%	
Eddie Collins	2B	99.99%	77.74%
Hank Aaron	RF	99.99%	
Lou Gehrig	1B	99.99%	special
Ted Williams	LF	99.99%	
Rogers Hornsby	2B	99.99%	78.11%
Willie Mays	CF	99.98%	
Jimmie Foxx	1B	99.98%	79.20%
Mel Ott	OF	99.97%	87.17%
Charlie Gehringer	2B	99.97%	85.03%
Stan Musial	OF	99.96%	93.24%
Frankie Frisch	2B	99.85%	84.47%
Al Simmons	OF	99.84%	75.38%

A Few Model Predictions



No Chance?

PLAYER	POS	Probability
Trot Nixon	OF	0.00081
Justin Morneau	1B	0.00216
Ryan Howard	1B	0.00267
Alfonso Soriano	OF	0.00294
David Ortiz	DH	0.00415
Mike Lowell	3B	0.00458
Adrian Beltre	3B	0.00731
Mark Teixeira	1B	0.00853
David Wright	3B	0.00831
Jason Varitek	C	0.00831
Jermaine Dye	OF	0.01789
Miguel Cabrera	1B	0.02177
Michael Young	3B	0.02882
Paul Konerko	1B	0.04166
Andruw Jones	OF	0.03128
Garrett Anderson	OF	0.06562
Chase Utley	2B	0.06665
Lance Berkman	1B	0.06714
Jim Edmonds	OF	0.08245
Carlos Lee	OF	0.09217
Victor Martinez	C	0.09458

Several too young...
HOWARD, MORNEAU,
WRIGHT, CABRERA,
UTLEY, MARTINEZ

Maybe???

PLAYER	POS	Probability
Carlos Delgado	1B	0.13326
Bobby Abreu	OF	0.1702
Jason Giambi	1B	0.24183
Miguel Tejada	SS	0.26106
Ichiro Suzuki	OF	0.31731
Edgar Renteria	SS	0.32892
Magglio Ordonez	OF	0.3697
Jason Kendall	C	0.38254
Johnny Damon	OF	0.43086
Moises Alou	OF	0.4411
Todd Helton	1B	0.47753

ICHIRO only 10 years...



Headed to
Cooperstown?

PLAYER	POS	Probability
Barry Bonds	OF	0.99986
Alex Rodriguez	3B	0.99725
Derek Jeter	SS	0.99268
Gary Sheffield	OF	0.98942
Mike Piazza	C	0.98673
Manny Ramirez	OF	0.97246
Ken Griffey, Jr.	OF	0.96469
Rafael Palmeiro	1B	0.95557
Chipper Jones	3B	0.95237
Frank Thomas	1B	0.94942
Vladimir Guerrero	OF	0.90668
Ivan Rodriguez	C	0.87593
Jeff Bagwell	1B	0.84358
Albert Pujols	1B	0.81416
Sammy Sosa	OF	0.80844
Jeff Kent	2B	0.78743
Craig Biggio	2B	0.73103
Omar Vizquel	SS	0.62581
Jim Thome	1B	0.6187
Bernie Williams	OF	0.58778

Some surprises here?

All-time Team

Imagine
this team
in Wrigley
Field!



DH - Aaron

Bench
SS - Ripken
CF - Mays
2B - Hornsby
1B - Foxx
3B - Traynor
OF - Ott
C - Piazza



All-time DODGER Team

“L.A.” (and/or sentimental) Picks

LF Kirk Gibson
CF Brett Butler
3B Pedro Guerrero
SS Maury Wills
1B Steve Garvey
C Steve Yeager

RF Reggie Smith
LF Dusty Baker
2B Jeff Kent
2B Davey Lopes
C Mike Scioscia



Ideas...



- Consider more complicated modeling
 - Predictive model approaches (ridge, LASSO etc)
 - Cross-validation
 - Consider other SABER metrics (OPS etc.)
 - Factor analysis to create new metrics for model
- Address other election criteria more explicitly:
 - DEFENSE
 - Contribution to team (metrics out there)
 - “Clutch” performance (post-season)
 - Awards (ROY, MVP, Gold Glove etc)
 - Character, Integrity, Sportsmanship (tougher!)

REFERENCES and SOURCES



- Hosmer, D. and Lemeshow, S. (2000), Applied Logistic Regression, Second Edition, Wiley, New York.
- National Baseball Hall of Fame and Museum
<http://baseballhall.org/>
- Baseball Almanac
www.baseball-almanac.com
- MLB.com
<http://mlb.mlb.com/>
- ESPN
www.espn.com
- RETROSHEET
www.retrosheet.org
- BASEBALL-REFERENCE.com
www.baseball-reference.com