Baseball III

Produced by Dr. Mario | UNC STOR 390

Linear Weights

- **Multiple Linear Regression**

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

Linear Weights

- **Baseball Application**
  - $Y = Runs\ for\ the\ Season$
  - $\vec{X} = [BB + HBP, S, D, T, HR, SB, CS]'$
  - $Y = \vec{X}'\vec{\beta} + \vec{\epsilon}$
  - $\hat{Y} = Predicted\ Runs$
  - $\hat{Y} = \vec{X}'\hat{\vec{\beta}}$

S = Single
D = Double
T = Triple
HR = Home Run
BB = Walk
HBP = Hit-by-Pitch
SB = Stolen Base
CS = Caught Stealing

Linear Weights

- Crude Estimation of Linear Weight for Home Run
  - $\widehat{\beta_{HR}} = E[\#\ of\ Runs|HR] = \frac{\#\ of\ Runs}{HR}$
  - Fact 1a: $\frac{4.8\ Runs\ Per\ Game}{38\ Batters\ Per\ Game} = 0.126\ Runs\ Per\ Batter$
  - Fact 2a: $\frac{4.8\ Runs\ Per\ Game}{13\ Batters\ Reach\ Base} = 0.369\ Runs\ Per\ Base\ Runner$
  - Suppose Batter Hits Home Run and Average of 1 Base Runner
  - Both Batter and Base Runner Score 100% of the Time
  - Fact 1b: $0.874\ Runs\ Per\ Home\ Run\ Batter$
  - Fact 2b: $0.631\ Runs\ Per\ Base\ Runner\ in\ a\ Home\ Run$
  - Therefore, $\frac{\#\ of\ Runs}{HR} = 0.874 + 0.631 = 1.505\ Runs$

Linear Weights

- Estimated Linear Weights Using Least Squares

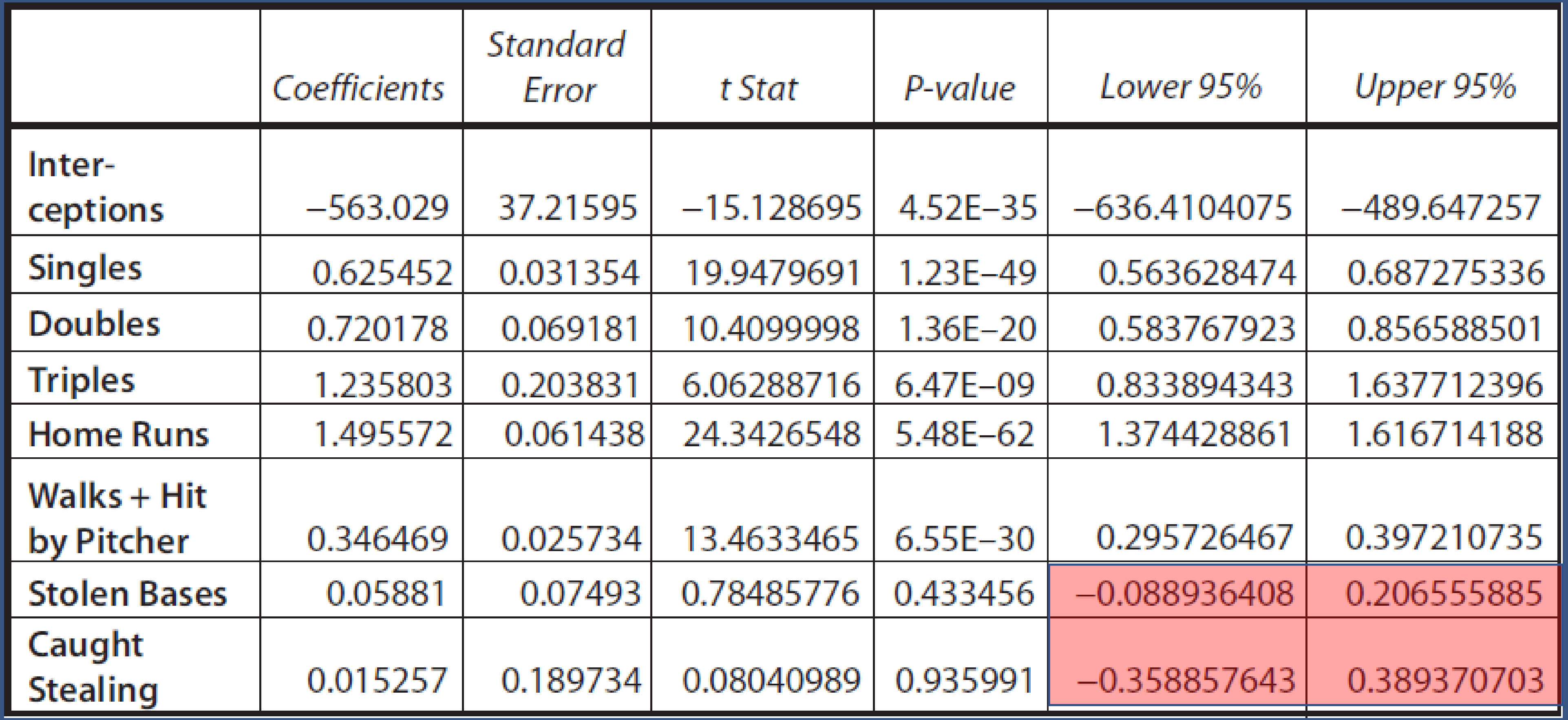
Predictor	Estimate
Constant	-563.03
Single	0.63
Double	0.72
Triple	1.24
HR	1.5
BB+HBP	0.35
SB	0.06
CS	0.02

$$n = 210$$
$$R^2 = 0.91$$
$$Adj.R^2 = 0.91$$

Doesn’t Add Marginal Value

Linear Weights

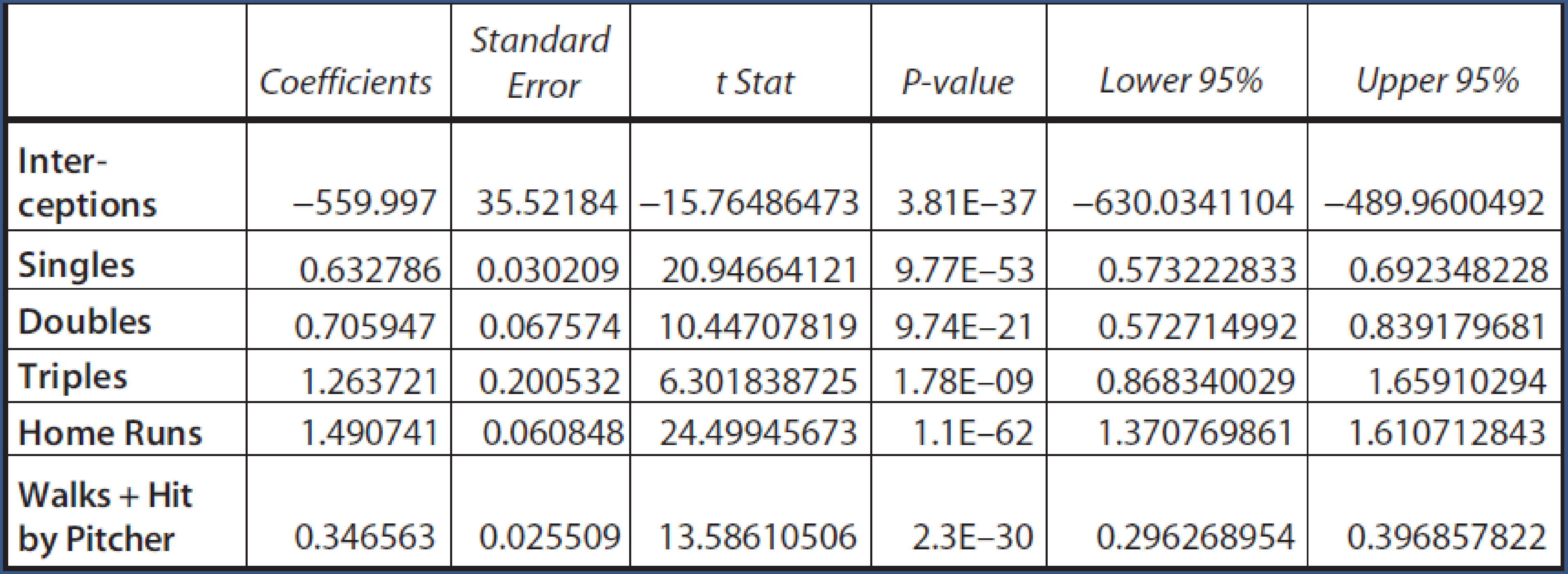
- Important Information From Linear Regression



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Interceptions	–563.029	37.21595	–15.128695	4.52E–35	–636.4104075	–489.647257
Singles	0.625452	0.031354	19.9479691	1.23E–49	0.563628474	0.687275336
Doubles	0.720178	0.069181	10.4099998	1.36E–20	0.583767923	0.856588501
Triples	1.235803	0.203831	6.06288716	6.47E–09	0.833894343	1.637712396
Home Runs	1.495572	0.061438	24.3426548	5.48E–62	1.374428861	1.616714188
Walks + Hit by Pitcher	0.346469	0.025734	13.4633465	6.55E–30	0.295726467	0.397210735
Stolen Bases	0.05881	0.07493	0.78485776	0.433456	–0.088936408	0.206555885
Caught Stealing	0.015257	0.189734	0.08040989	0.935991	–0.358857643	0.389370703

Linear Weights

- Important Information From Linear Regression
  - Removal of Insignificant Variables

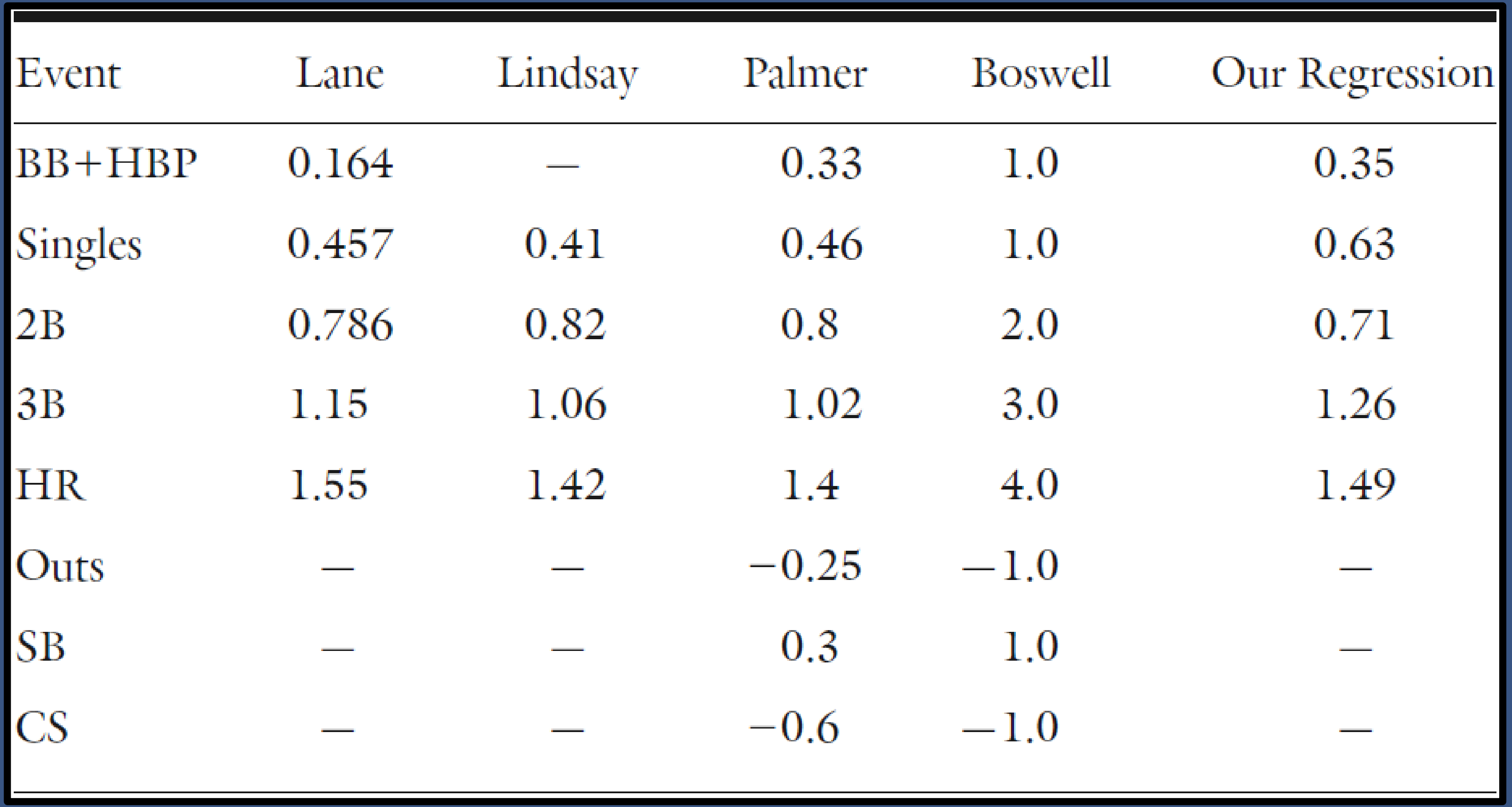


	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Interceptions	–559.997	35.52184	–15.76486473	3.81E–37	–630.0341104	–489.9600492
Singles	0.632786	0.030209	20.94664121	9.77E–53	0.573222833	0.692348228
Doubles	0.705947	0.067574	10.44707819	9.74E–21	0.572714992	0.839179681
Triples	1.263721	0.200532	6.301838725	1.78E–09	0.868340029	1.65910294
Home Runs	1.490741	0.060848	24.49945673	1.1E–62	1.370769861	1.610712843
Walks + Hit by Pitcher	0.346563	0.025509	13.58610506	2.3E–30	0.296268954	0.396857822

- $RMSE = 210$ and $MAD = 210$ **(Outperforms Previous)**

Linear Weights

- **Historical Progression**



	1916	1950-1960	1978	1989	
Event	Lane	Lindsay	Palmer	Boswell	Our Regression
BB+HBP	0.164	—	0.33	1.0	0.35
Singles	0.457	0.41	0.46	1.0	0.63
2B	0.786	0.82	0.8	2.0	0.71
3B	1.15	1.06	1.02	3.0	1.26
HR	1.55	1.42	1.4	4.0	1.49
Outs	—	—	−0.25	−1.0	—
SB	—	—	0.3	1.0	—
CS	—	—	−0.6	−1.0	—

Linear Weights

- Evaluation of Hitters
  - Imagine if Team Had Only Barry Bonds (2004)
  - Approximately,

$$26.72 \times 162 = 4329\ Outs\ Per\ Season$$

  - Bonds Hit 45 HR and Had 240.29 Outs
  - Therefore, Bonds Hit

$$\frac{45}{240.29}\ Home\ Runs\ Per\ Out$$

  - Scaling Up, We Expect a Team of Bonds to Hit

$$4329 \times \frac{45}{240.29} = 811\ Home\ Runs\ Per\ Season$$

  - Using Linear Weights, We Expect 3,259 Runs Per Season which Can Be Thought of 20.12 Runs Per Game

Linear Weights

- OBP, SLG, OPS, and Runs Created
  - *Moneyball* Highlights the Importance of OBP
  - From 2000-2006, Average OBP was 33%
  - Purpose of OPS = Value Power Hitters
  - Recall:

$$\begin{aligned} OPS &= OBP + SLG \\ &= 1 \times OBP + 1 \times SLG \end{aligned}$$

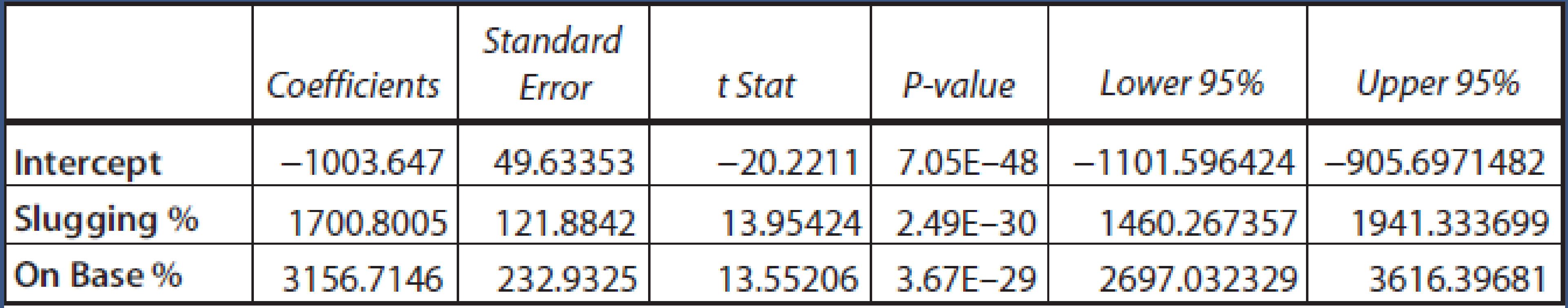
Equal Weights

  - Which Covariate (OBP or SLG) is Better for Predicting Runs?

Linear Weights

- OBP, SLG, OPS, and Runs Created
  - Multiple Regression

$$Runs = \beta_0 + \beta_1(SLG) + \beta_2(OBP) + \epsilon$$



	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	–1003.647	49.63353	–20.2211	7.05E–48	–1101.596424	–905.6971482
Slugging %	1700.8005	121.8842	13.95424	2.49E–30	1460.267357	1941.333699
On Base %	3156.7146	232.9325	13.55206	3.67E–29	2697.032329	3616.39681

$$n = 180 \;\&\; R^2 = 0.91 \;\&\; Adj.R^2 = 0.91$$

- Summary: OBP Twice as Valuable as SLG

Linear Weights

- Runs Created Above Average
  - How Many More Runs if Average Team Added a Player?
  - Average Team (2000-2006) Versus Ichiro (2004)

Hit Type	Average Team	Ichiro 2004
Single	972.08	225
Double	296	24
Triple	30.82	5
HR	177.48	8
BB+HBP	599.88	60
Outs	4329	451

Linear Weights

- Runs Created Above Average
  - If Added, Rest of Players Will Cost an Approximate

    $$4329 - 451 = 3878\ Outs$$

  - For the Rest of The Team, This is Equivalent to

    $$\frac{3878}{4329} = 88\%\ of\ Total\ Outs$$

  - Singles With Ichiro Added to Roster

    $$Singles = 0.88(Singles\ of\ Team) + (Singles\ of\ Ichiro)$$

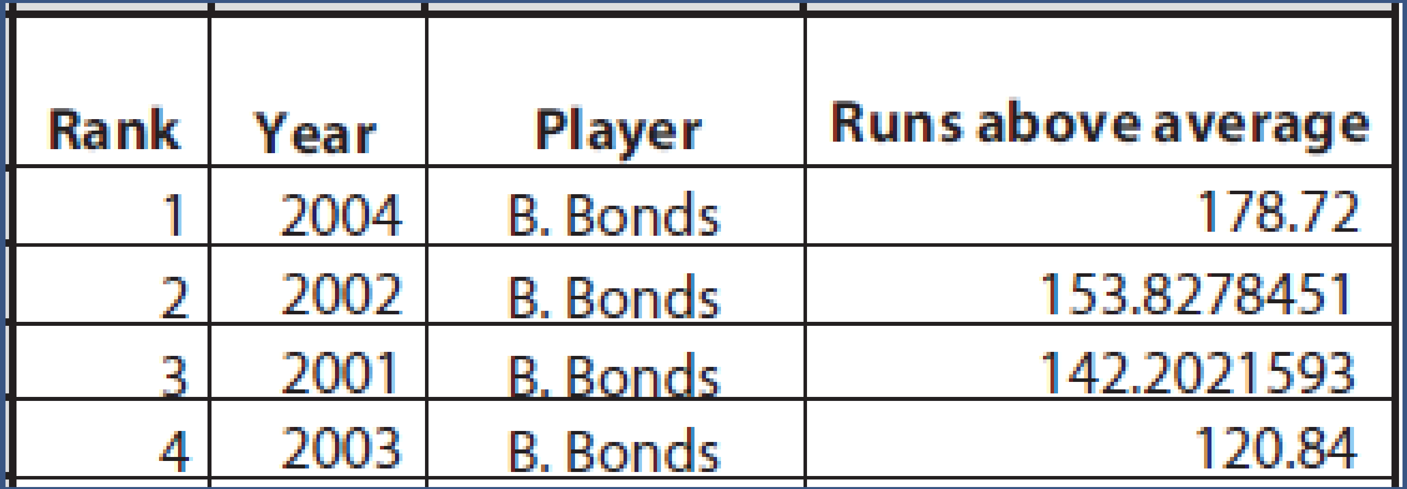
Linear Weights

- Runs Created Above Average

Hit Type	Average Team	Ichiro	Ichiro+Team
Single	972.08	225	1095.73
Double	296	24	289.13
Triple	30.82	5	32.60
HR	177.48	8	166.98
BB+HBP	599.88	60	597.33

Linear Weights

- Runs Created Above Average
  - Predicted Runs of Average Team = 780
  - Predicted Runs of Ichiro+Average Team = 839
  - Added Value of Ichiro = 839-780 = 59 Runs Above Average
  - Perspective:



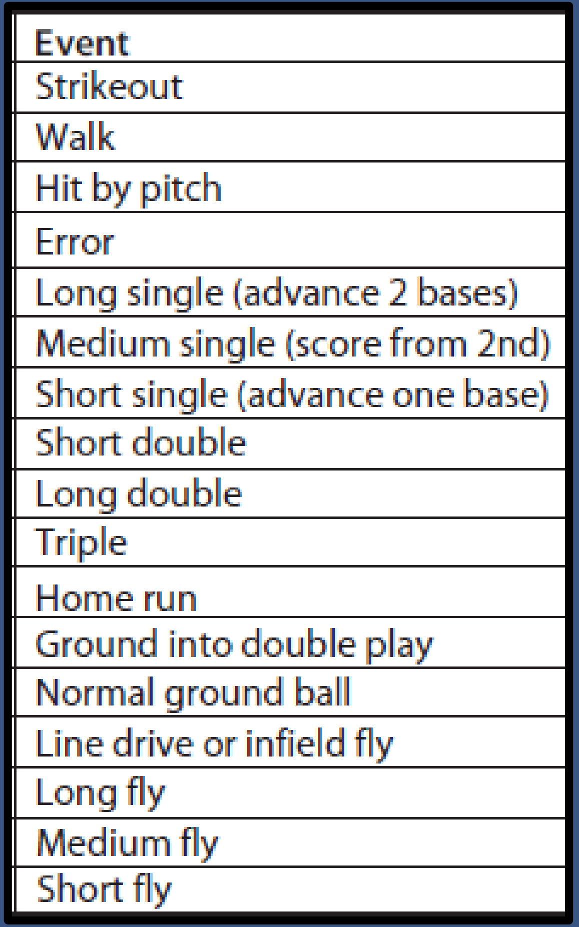
Rank	Year	Player	Runs above average
1	2004	B. Bonds	178.72
2	2002	B. Bonds	153.8278451
3	2001	B. Bonds	142.2021593
4	2003	B. Bonds	120.84

Monte Carlo Simulation

- Recall Evaluation of Hitter Effectiveness
  - Runs Created
  - Linear Weights
  - Both Based on Team Data
  - Scaled Player Information for Prediction
- Problem: Player Hits HR 50% of Time = 54 RC/G
- Definition of Monte Carlo Simulation
  - Developing a Computer Model to Repeatedly Play Out an Uncertain Situation
  - Used Across All Industries
  - Term Coined by Polish Physicist Stanislaw Ulam
  - Simple Simulation Shows Previously Discussed Player = 27 RC/G

Monte Carlo Simulation

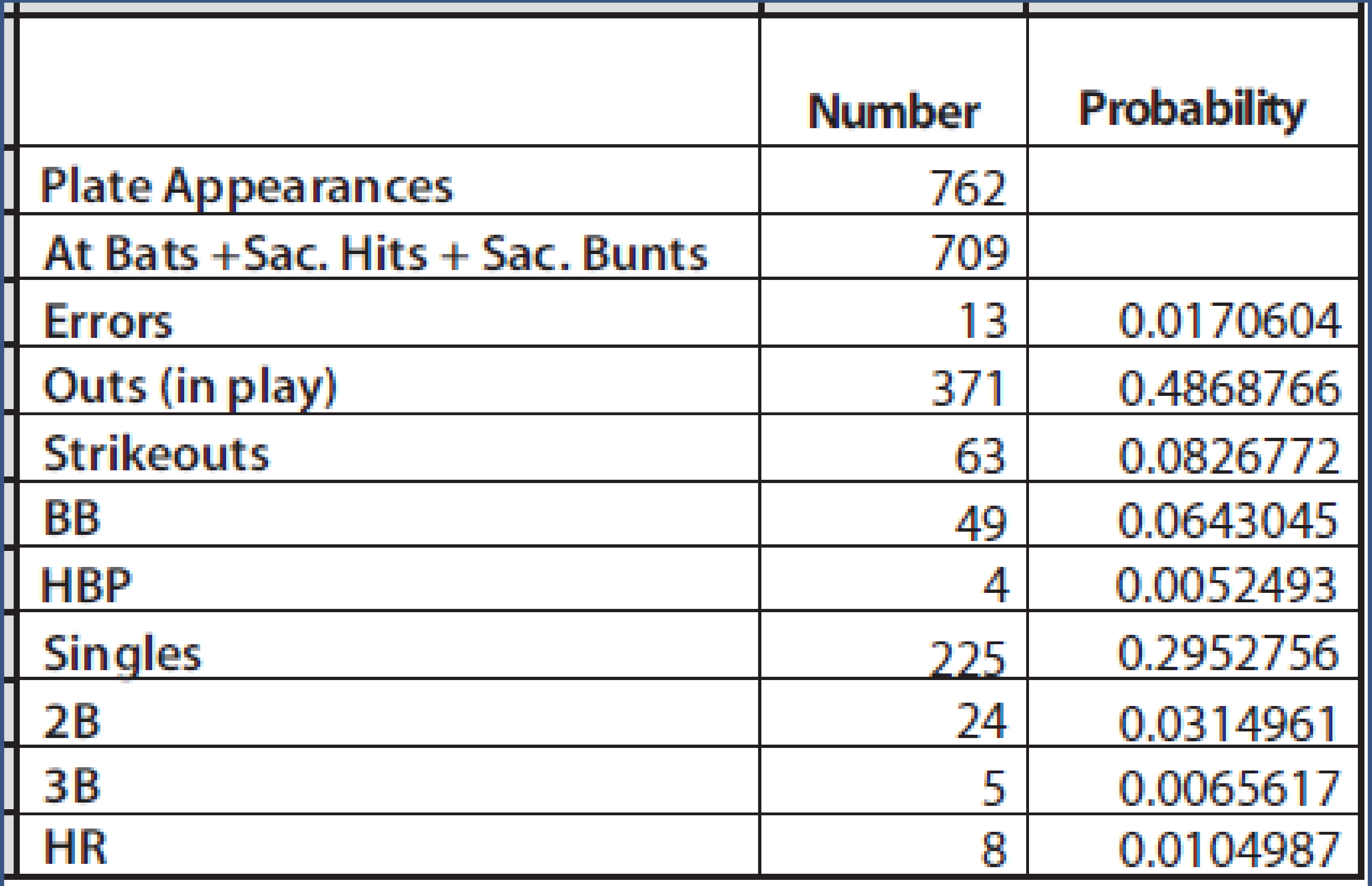
- Simulating Runs from Team Full of Ichiros
  - Possible Plate Appearances Events
  - Long List of Assumptions
    - Errors Advance All Base Runners 1 Base
    - Long Single Advances Each Runner 2 Bases
    - Short Single Advances All Runners 1 Base
    - Short Double Advances Each Runner 2 Bases
    - Long Double Scores a Runner from First
    - Etc.
  - Assign Probabilities According to Relative Frequencies of Player
  - Program for Simulation



Event
Strikeout
Walk
Hit by pitch
Error
Long single (advance 2 bases)
Medium single (score from 2nd)
Short single (advance one base)
Short double
Long double
Triple
Home run
Ground into double play
Normal ground ball
Line drive or infield fly
Long fly
Medium fly
Short fly

Monte Carlo Simulation

- Simulating Runs from Team Full of Ichiros
  - Probabilities Based on Ichiro 2004 Statistics



	Number	Probability
Plate Appearances	762	
At Bats +Sac. Hits + Sac. Bunts	709	
Errors	13	0.0170604
Outs (in play)	371	0.4868766
Strikeouts	63	0.0826772
BB	49	0.0643045
HBP	4	0.0052493
Singles	225	0.2952756
2B	24	0.0314961
3B	5	0.0065617
HR	8	0.0104987

Monte Carlo Simulation

- Simulating Runs from Team Full of Ichiros
  - Probabilities of Special Cases
    - 30% of Singles are Long Singles
    - 50% of Singles are Medium Singles
    - 20% of Singles are Short Singles
    - 53.8% of Outs in Play are Ground Balls
    - 15.3% of Outs in Play are Infield Flies
    - 30.9% of Outs in Play are Fly Balls
    - Etc.
  - Result of Simulation = Within 1% of True Actual Runs Per Game
  - Specific to Ichiro
    - Random Number < 0.295 = Single
    - Random Number < 0.487 = Out
  - Goal of Simulation
    - Estimate # of Runs for Thousands of Innings
    - Average Across All Innings
    - Multiply by $\frac{26.72}{3} \approx 9$ to estimate RC/G

Monte Carlo Simulation

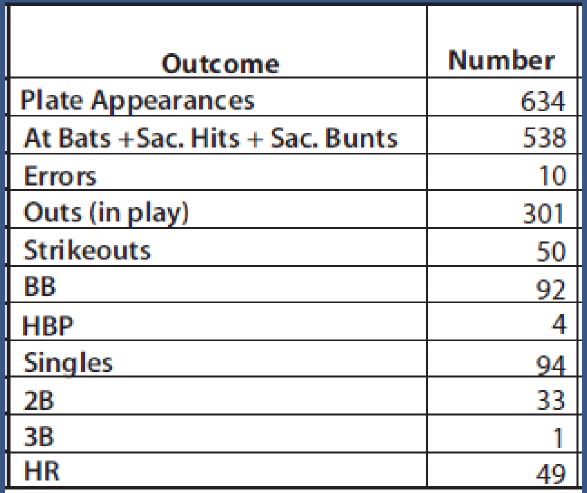
- **Results Under Simulation**

Player	Year	RC/G
Ichiro	2004	6.92
Nomar	1997	5.91
Bonds	0.72	21.02

Problem: Unusual # of Intentional Walks
Eliminating Intentional Walks: 15.98 RC/G

Monte Carlo Simulation

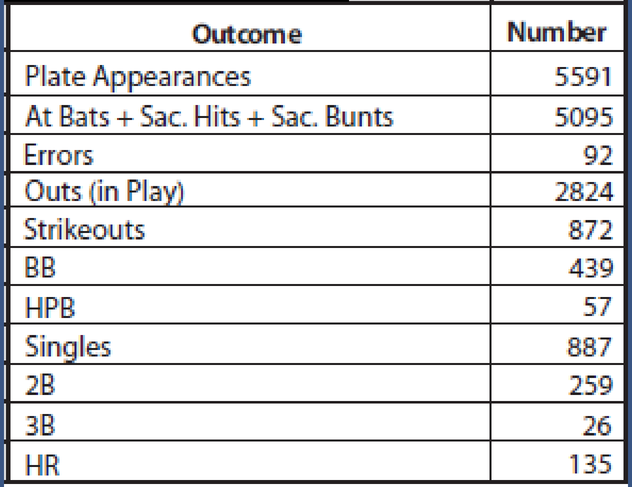
- Added Value of Albert Pujols Measured by Runs



Outcome	Number
Plate Appearances	634
At Bats +Sac. Hits + Sac. Bunts	538
Errors	10
Outs (in play)	301
Strikeouts	50
BB	92
HBP	4
Singles	94
2B	33
3B	1
HR	49

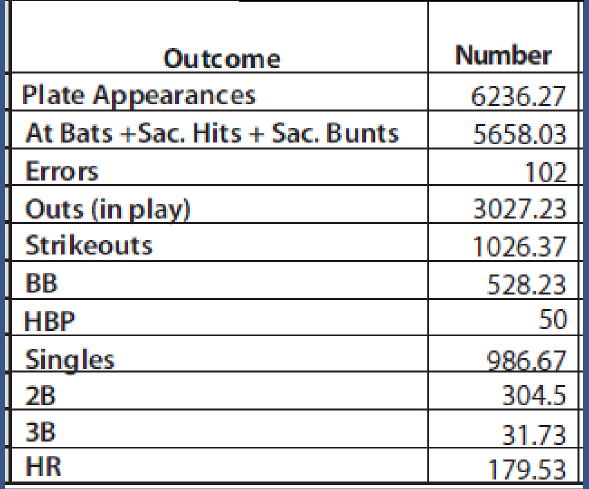
Pujols Alone

Team Without



Outcome	Number
Plate Appearances	5591
At Bats + Sac. Hits + Sac. Bunts	5095
Errors	92
Outs (in Play)	2824
Strikeouts	872
BB	439
HPB	57
Singles	887
2B	259
3B	26
HR	135

Average Team



Outcome	Number
Plate Appearances	6236.27
At Bats +Sac. Hits + Sac. Bunts	5658.03
Errors	102
Outs (in play)	3027.23
Strikeouts	1026.37
BB	528.23
HBP	50
Singles	986.67
2B	304.5
3B	31.73
HR	179.53

America’s Greatest Pastime

Final Inspiration

If you don’t like sports,
you may like baseball.

- Mahatma Mario