

B8899-001: Applied Regression Analysis (Juran)

Term Project:

MLB Free Agent Salaries



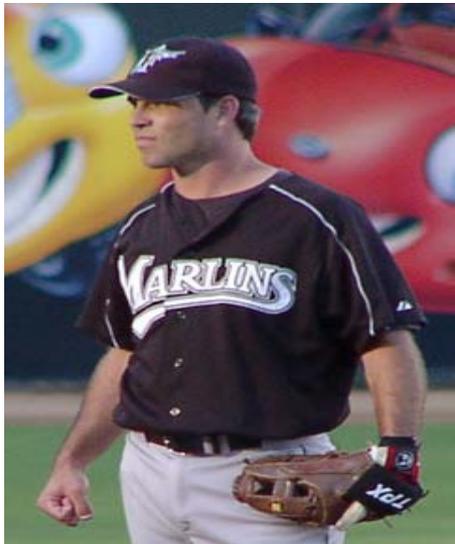
Hong Moo (Henry) Jun

Calvin Kuo

Mark McMillan

Shiraz Rehman

October 18, 2005



The Management Problem

- Can we predict a Major League Baseball (“MLB”) Free Agent (“FA”) **hitter's** annualized salary from his previous 3 years’ performance?
- We regressed the annual salaries of 217 free agents signed for the 2004 and 2005 MLB seasons, against performance data from each of their previous 3 seasons

Challenges

- Cleaning and formatting data
 - Lots of Excel vlookups
 - Data mining/cleansing/formatting took approximately 70% of the total project time
- Standardization of data variables (player names, team codes) across multiple data source sites (espn.com, mlb.com, statspass.com)
- Relevance of the model due to outside intangibles that General Managers take into consideration

Variables

- Independent: Annual salary
- Dependents: 9 dummy, 15 regular; 24 total
 - Physical attributes: age at signing; height; weight
 - Performance attributes:
 - 3-year, weighted-average batting statistics
 - 50% weight given to year t-1; 30% t-2; 20% t-3
 - AB, H, 2B, HR, R, RBI, K, BB, SB, AVG, OBP, SLG
 - » Variables also weighted by AB to remove injuries and lack of data for rookies (the variable AB itself contains this info)
 - Fielding statistics
 - Defensive fielding percentage: FPCT
 - Team-based statistics
 - Total wins for former team
 - Team-switching factor (aka “hometown discount”)
 - Other: Position, Inflation factor, Contract length

Correlation/Collinearity

	Player Salary	Btwn 03/04	Contract Length	Same Team	DH	C	1B	2B	SS	3B	Age	Ht	Wt	AB	R	2B	HR	RBI	BB	SB	AVG	OBP	FPCT	SLG	Team Wins
Player Salary	1.00	-0.17	0.83	-0.11	-0.03	-0.13	0.09	-0.13	0.05	0.07	-0.21	0.11	0.01	0.60	0.71	0.62	0.70	0.74	0.59	0.35	0.50	0.51	-0.02	0.63	0.16
Btwn 03/04		1.00	-0.16	0.22	0.06	-0.04	0.00	-0.04	-0.12	0.07	-0.13	-0.06	0.06	-0.03	-0.02	-0.01	-0.02	-0.04	0.00	0.06	0.01	0.02	-0.05	-0.02	-0.06
Contract Length			1.00	-0.18	-0.09	-0.04	0.04	-0.10	0.04	0.04	-0.22	0.08	-0.06	0.44	0.54	0.46	0.47	0.51	0.44	0.30	0.39	0.40	0.03	0.44	0.16
Same Team				1.00	0.01	0.10	-0.04	0.01	-0.04	0.01	-0.13	-0.03	-0.10	-0.20	-0.17	-0.12	-0.22	-0.20	-0.12	0.00	0.06	0.06	-0.07	-0.09	0.09
DH					1.00	-0.08	-0.05	-0.07	-0.06	-0.06	0.16	0.01	0.19	-0.02	0.01	0.00	0.11	0.09	0.11	-0.08	0.05	0.12	-0.33	0.12	0.05
C						1.00	-0.14	-0.20	-0.17	-0.15	0.12	0.10	0.14	-0.36	-0.44	-0.32	-0.26	-0.31	-0.36	-0.28	-0.23	-0.26	0.13	-0.21	0.03
1B							1.00	-0.13	-0.11	-0.10	0.11	0.37	0.20	0.03	0.07	0.09	0.20	0.19	0.28	-0.19	0.07	0.26	0.10	0.20	0.05
2B								1.00	-0.16	-0.14	-0.08	-0.24	-0.27	0.07	0.03	-0.02	-0.25	-0.18	-0.07	0.15	0.04	-0.05	-0.04	-0.25	-0.03
SS									1.00	-0.12	-0.08	-0.20	-0.30	0.22	0.15	0.18	-0.08	0.03	-0.02	0.11	-0.02	-0.18	-0.07	-0.16	0.05
3B										1.00	-0.06	0.00	-0.01	0.14	0.06	0.09	0.15	0.17	0.06	-0.09	-0.14	-0.11	-0.07	0.03	0.01
Age											1.00	0.03	-0.03	-0.11	-0.16	-0.17	-0.06	-0.08	0.00	-0.18	0.02	0.08	-0.11	-0.04	0.01
Height												1.00	0.65	-0.10	-0.08	-0.01	0.22	0.15	0.10	-0.26	-0.02	0.12	0.08	0.29	0.06
Weight													1.00	-0.20	-0.13	-0.04	0.27	0.15	0.08	-0.35	-0.12	0.13	0.03	0.34	-0.08
AB														1.00	0.90	0.84	0.60	0.78	0.65	0.47	0.41	0.29	-0.08	0.36	0.00
R															1.00	0.84	0.71	0.83	0.77	0.54	0.51	0.51	-0.06	0.55	0.04
2B																1.00	0.63	0.80	0.63	0.34	0.49	0.41	-0.08	0.52	-0.01
HR																	1.00	0.92	0.68	0.08	0.31	0.45	-0.10	0.84	0.00
RBI																		1.00	0.73	0.16	0.43	0.46	-0.12	0.74	0.01
BB																			1.00	0.35	0.25	0.67	-0.16	0.50	0.00
SB																				1.00	0.26	0.23	0.05	0.02	0.05
AVG																					1.00	0.66	-0.08	0.54	0.06
OBP																						1.00	-0.14	0.62	0.04
FPCT																							1.00	-0.10	-0.07
SLG																								1.00	0.06
Team Wins																									1.00

- Contract length, ABs, Rs, 2Bs, HRs, RBIs, and SLG have a high correlation to Player Salary
- BB and RBI each have high correlation to ABs, Rs, 2Bs, and HRs

Regression Output: Full model

Predictor	Coef	SE Coef	T	P
Constant	-18490266	6410418	-2.88	0.004
Between 03/04	-1017191	255461	-3.98	0.000
Same_Team	-211930	253654	-0.84	0.404
DH	-473551	782281	-0.61	0.546
C	2051672	385610	5.32	0.000
1B	-101795	515116	-0.20	0.844
2B	370667	401160	0.92	0.357
SS	1118722	453496	2.47	0.015
3B	925232	459100	2.02	0.045
Age	-168645	34224	-4.93	0.000
Height	140709	87462	1.61	0.109
Weight	-18120	9764	-1.86	0.065
AB	-4402	2660	-1.66	0.100
R	1750	19185	0.09	0.927
Doubles	-20198	30514	-0.66	0.509
HR	159025	59192	2.69	0.008
RBI	52624	20513	2.57	0.011
BB	10985	20209	0.54	0.587
SB	92287	21691	4.25	0.000
AVG	38084661	13649281	2.79	0.006
OBP	15364190	11652419	1.32	0.189
FPCT	1916746	2767686	0.69	0.489
SLG	-7220570	6105668	-1.18	0.238
Team_Wins	28473	10171	2.80	0.006

S = 1654226	R-Sq = 76.9%	R-Sq(adj) = 74.1%
-------------	--------------	-------------------

Potential Pitfalls

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.907
R Square	0.822
Adjusted R Square	0.821
Standard Error	1,377,475
Observations	217

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	1.88E+15	9.39E+14	495	0.000
Residual	214	4.06E+14	1.90E+12		
Total	216	2.28E+15			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	(2,892,836)	219,537	(13.18)	0.000	(3,325,567)	(2,460,105)
Contract Length	1,759,905	96,891	18.16	0.000	1,568,921	1,950,888
RBI	55,149	4,337	12.71	0.000	46,599	63,698

- This model has a terrific R^2 , but contract length does not assist a general manager in determining the fair value of a free agent.
- Shows that general managers sign players to long contracts if they're great players (obvious!).
- Contract length is a predictive variable, but not in any way that would help GMs, so we've removed it from our model.

Best Subsets

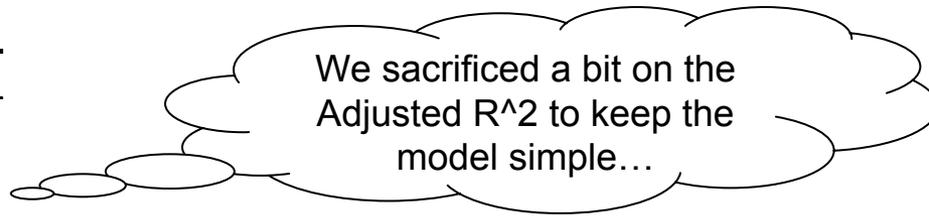
Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	a	D	1	2	3	g	h	h	A	B	R	H	B	B	S	V	B	C	L	n
1	54.8	54.6	164.1	2190947																				
1	51.0	50.8	195.6	2280572										X										
2	60.3	59.9	120.5	2058735													X		X					
2	59.4	59.0	128.3	2082826										X										X
3	63.1	62.6	98.8	1988428										X										X
3	62.9	62.4	100.5	1994119															X		X			
4	65.8	65.1	78.7	1920247	X												X			X	X			
4	65.7	65.0	79.6	1923378													X			X		X		
5	68.6	67.8	57.3	1844548	X							X					X			X	X			
5	68.2	67.4	60.7	1856341								X					X			X		X		
6	71.3	70.5	36.4	1766106	X	X								X					X	X	X			
6	71.2	70.4	37.0	1768564	X	X											X		X	X	X			
7	72.6	71.7	27.6	1730022	X	X											X		X		X			X
7	72.6	71.7	27.7	1730476	X	X											X		X		X			X
8	73.8	72.8	20.0	1697294	X	X											X	X	X	X	X			X
8	73.8	72.8	20.0	1697296	X	X											X	X	X	X	X			X
9	74.4	73.3	16.6	1680487	X	X											X	X	X	X	X			X
9	74.4	73.2	17.1	1682237	X	X											X	X	X	X	X			X
10	75.0	73.8	13.9	1665884	X	X			X	X	X						X		X	X	X			X
10	74.9	73.7	14.6	1668656	X	X					X	X					X	X	X	X	X			X
11	75.3	74.0	13.2	1659238	X	X				X	X	X					X	X	X	X	X			X
11	75.2	73.9	13.8	1661372	X	X					X	X	X				X	X	X	X	X			X
12	75.5	74.1	13.1	1654592	X	X			X	X	X						X	X	X	X	X			X
12	75.5	74.1	13.1	1654789	X	X			X	X	X	X					X	X	X	X	X			X
13	75.9	74.3	12.2	1646938	X	X			X	X	X	X	X				X	X	X	X	X			X
13	75.8	74.3	12.7	1648889	X	X			X	X	X	X	X				X	X	X	X	X			X
14	76.2	74.5	12.0	1642053	X	X			X	X	X	X	X				X	X	X	X	X			X
14	76.1	74.5	12.1	1642246	X	X			X	X	X	X	X	X			X	X	X	X	X			X
15	76.5	74.7	11.5	1635515	X	X			X	X	X	X	X	X			X	X	X	X	X			X
15	76.3	74.5	13.0	1642017	X	X			X	X	X	X	X	X			X	X	X	X	X			X
16	76.6	74.7	12.6	1635860	X	X			X	X	X	X	X	X			X	X	X	X	X			X
16	76.6	74.7	12.7	1636297	X	X			X	X	X	X	X	X			X	X	X	X	X			X
17	76.7	74.7	13.8	1636547	X	X			X	X	X	X	X	X			X	X	X	X	X			X
17	76.7	74.7	13.9	1636976	X	X			X	X	X	X	X	X			X	X	X	X	X			X
18	76.7	74.6	15.1	1637800	X	X			X	X	X	X	X	X			X	X	X	X	X			X
18	76.7	74.6	15.2	1638265	X	X			X	X	X	X	X	X			X	X	X	X	X			X
19	76.8	74.6	16.7	1640320	X	X			X	X	X	X	X	X			X	X	X	X	X			X
19	76.8	74.5	16.8	1640750	X	X			X	X	X	X	X	X			X	X	X	X	X			X
20	76.8	74.5	18.4	1643201	X	X			X	X	X	X	X	X			X	X	X	X	X			X
20	76.8	74.5	18.4	1643207	X	X			X	X	X	X	X	X			X	X	X	X	X			X
21	76.9	74.4	20.1	1645941	X	X			X	X	X	X	X	X			X	X	X	X	X			X
21	76.8	74.3	20.3	1647006	X	X			X	X	X	X	X	X	X			X	X	X	X			X
22	76.9	74.3	22.0	1649992	X	X			X	X	X	X	X	X	X			X	X	X	X			X
22	76.9	74.2	22.0	1650124	X	X			X	X	X	X	X	X	X			X	X	X	X			X
23	76.9	74.1	24.0	1654226	X	X			X	X	X	X	X	X	X			X	X	X	X			X

T
a
m
e
e
e
i
i
2
F
W
A
g
g
B
R
A
O
P
S
i
n
s
H
B
B
S
V
B
C
L
n
s
R
I
B
B
G
P
T
G
s
1
R
I
B
B
G
P
T
G
s

Regression Output: Reduced Model

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.845
R Square	0.713
Adjusted R Square	0.705
Standard Error	1,766,106
Observations	217



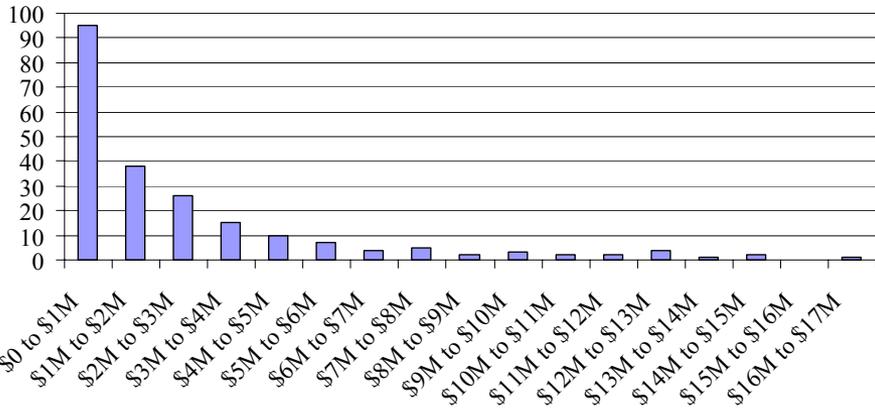
ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	6	1.629E+15	2.715E+14	87.044	0.000
Residual	210	6.550E+14	3.119E+12		
Total	216	2.284E+15			

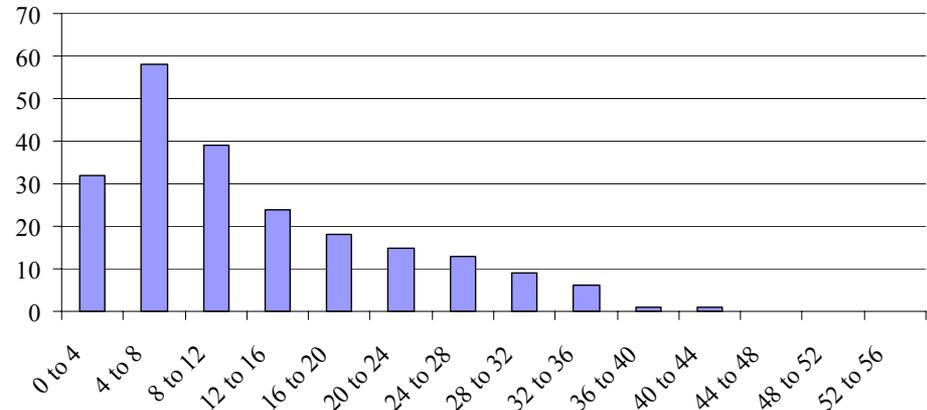
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	(6,527,705)	1,843,281	(3.541)	0.0005	(10,161,409)	(2,894,000)
Between '03/'04	(1,285,833)	256,320	(5.017)	0.0000	(1,791,123)	(780,542)
C	1,551,313	345,509	4.490	0.0000	870,202	2,232,424
Age	(158,876)	32,954	(4.821)	0.0000	(223,839)	(93,912)
HR	223,723	14,321	15.622	0.0000	195,492	251,955
SB	97,839	15,720	6.224	0.0000	66,851	128,828
AVG	43,480,437	6,061,013	7.174	0.0000	31,532,212	55,428,662

Histograms of Selected Variables

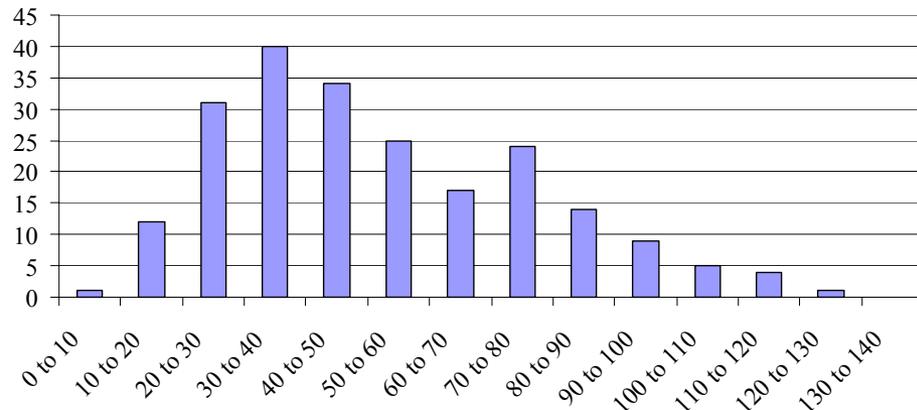
Histogram of Player Salaries



Histogram of Home Runs

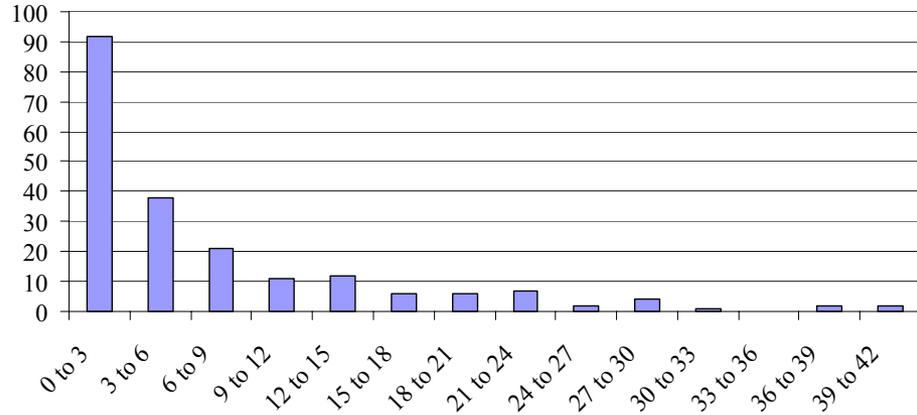


Histogram of RBIs

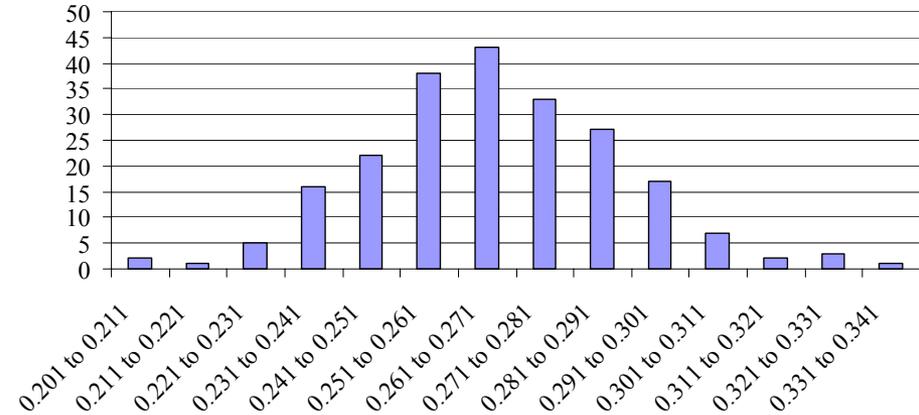


Histograms of Selected Variables

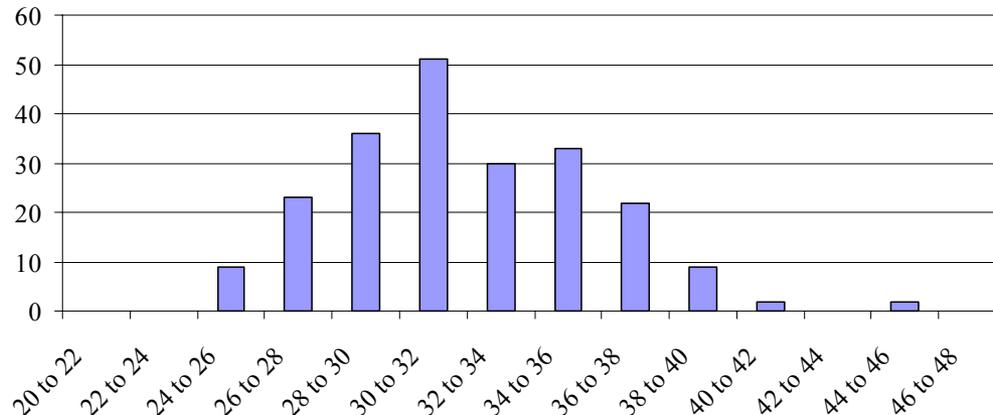
Histogram of Stolen Bases



Histogram of Batting Average

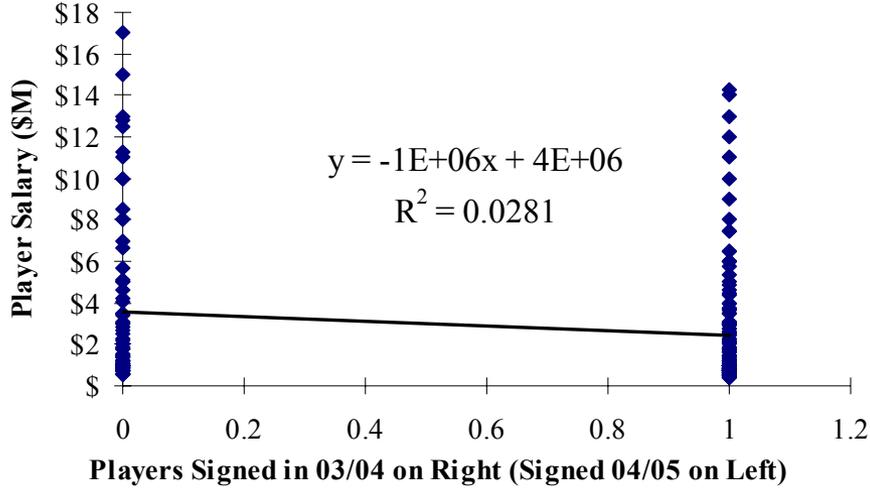


Histogram of Ages

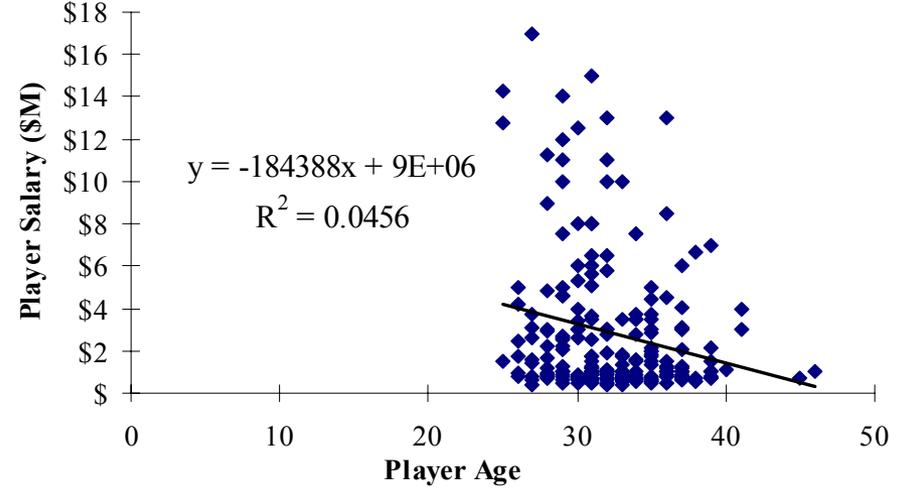


Scatter Plots of Selected Variables

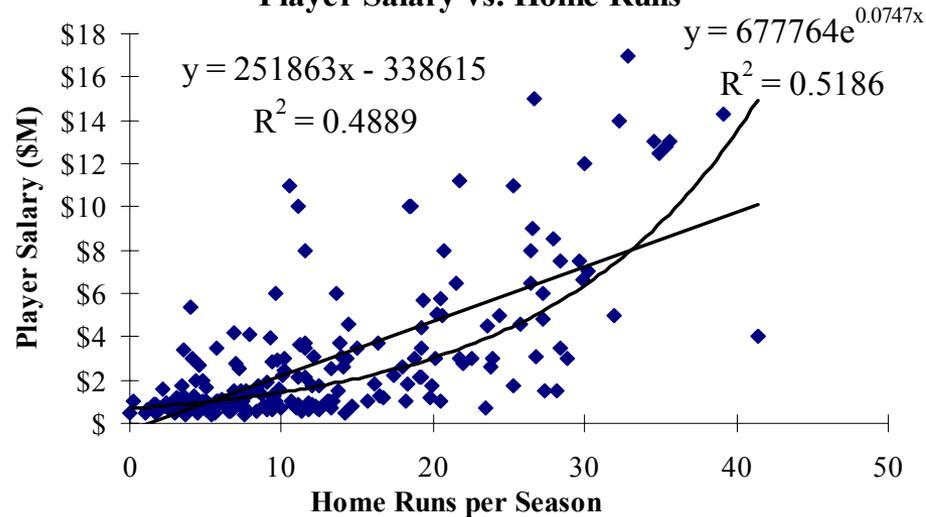
Player Salary vs. Signing Year



Player Salary vs. Age

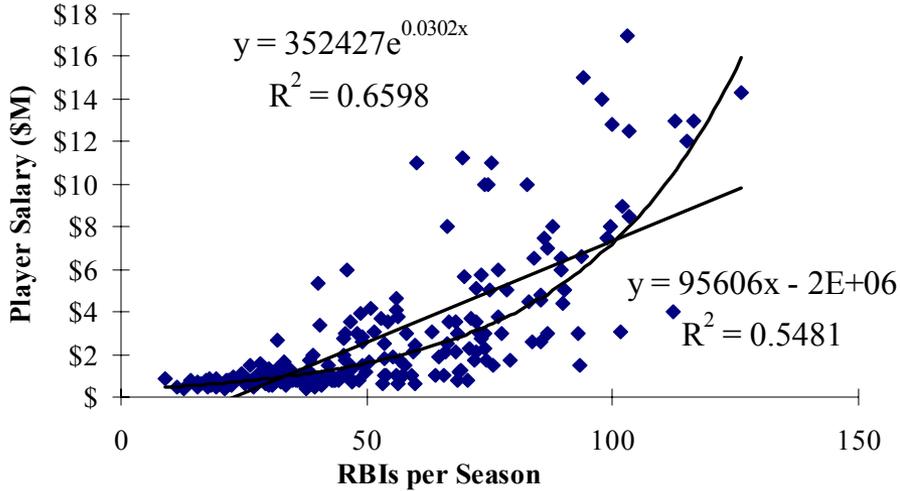


Player Salary vs. Home Runs

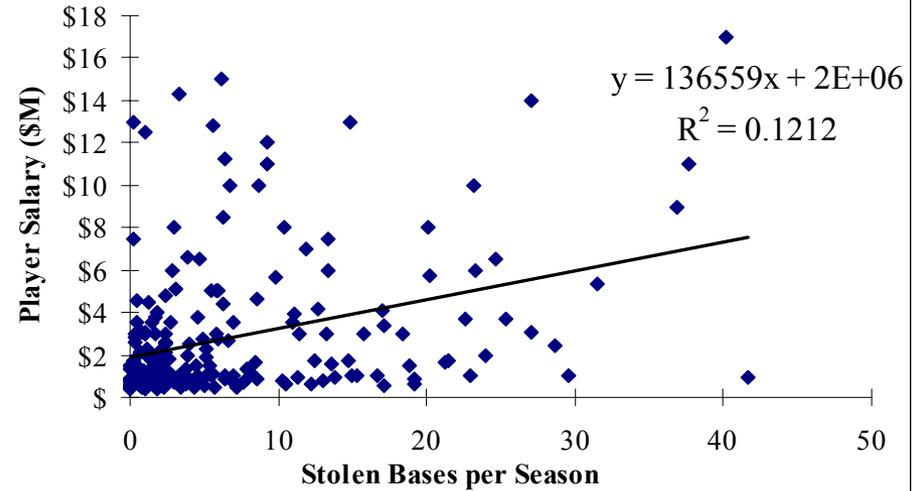


Scatter Plots of Selected Variables

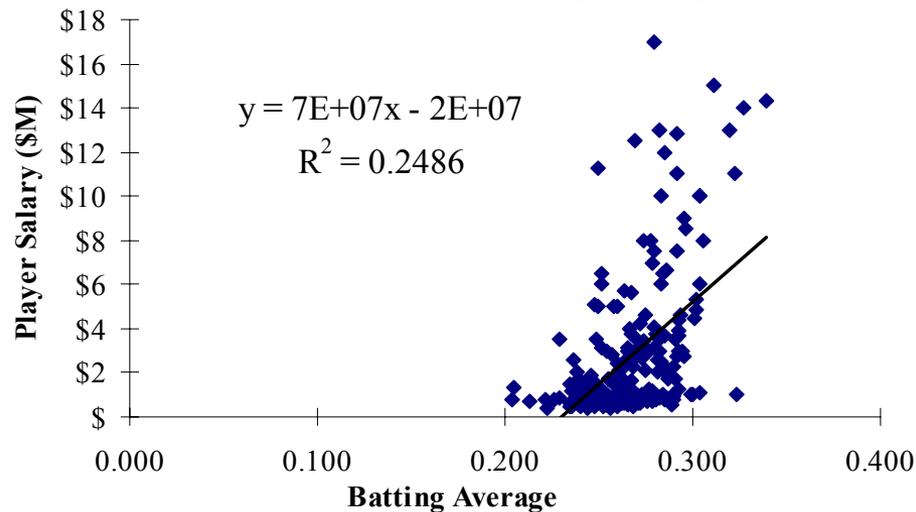
Player Salary vs. RBIs



Player Salary vs. Stolen Bases

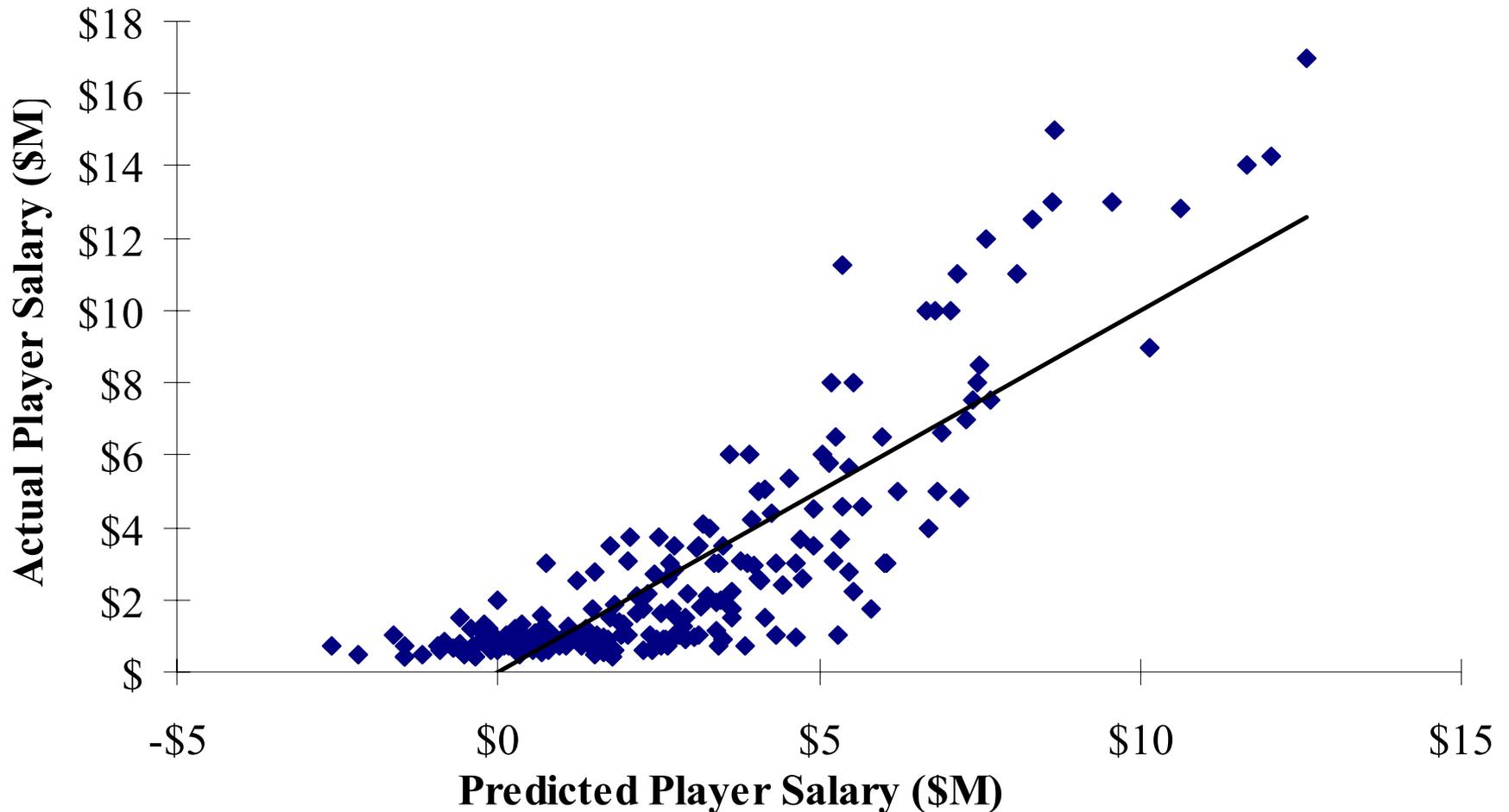


Player Salary vs. Batting Average

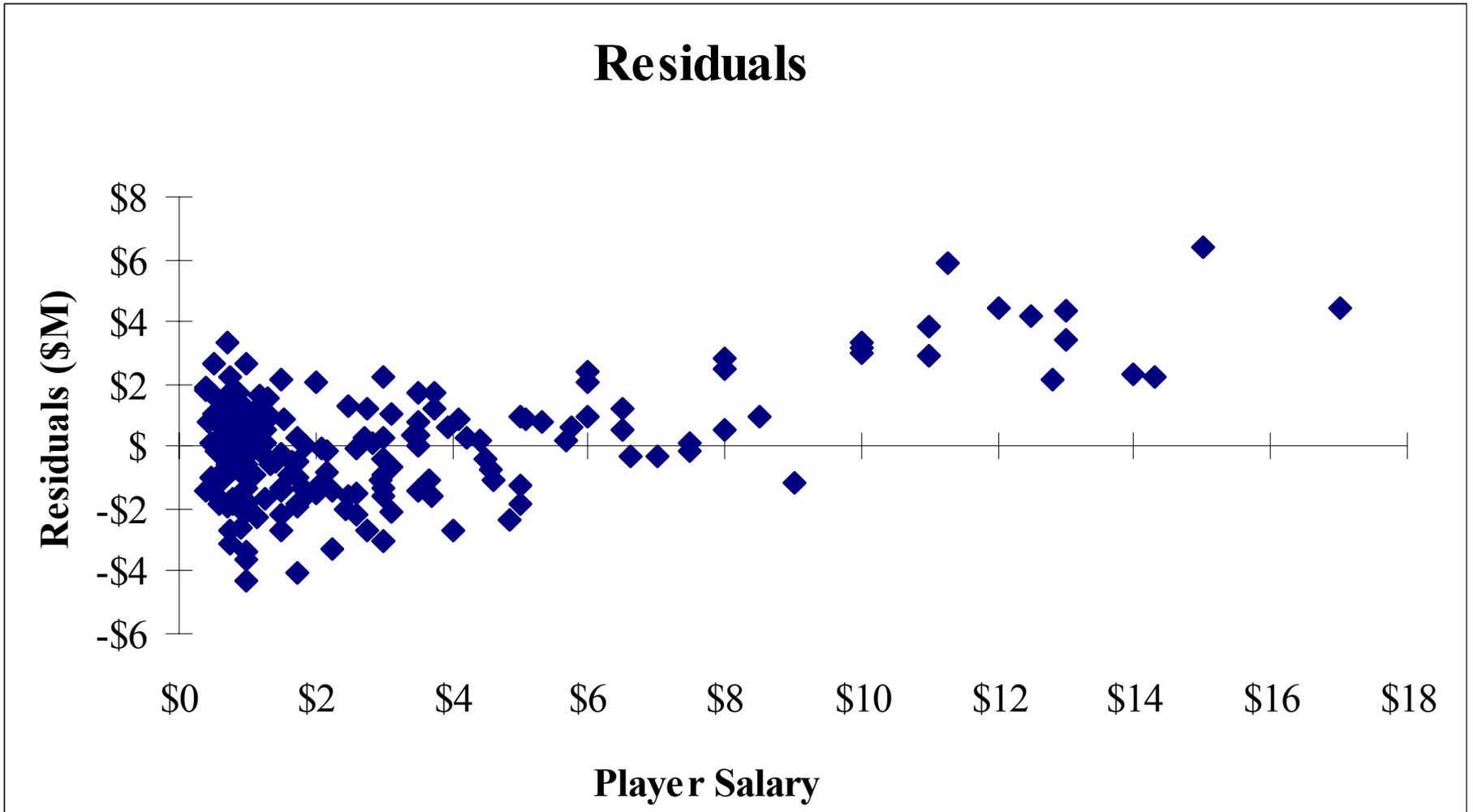


Residuals Analysis

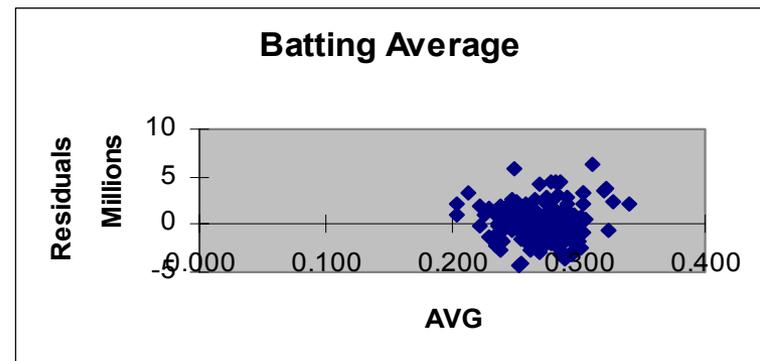
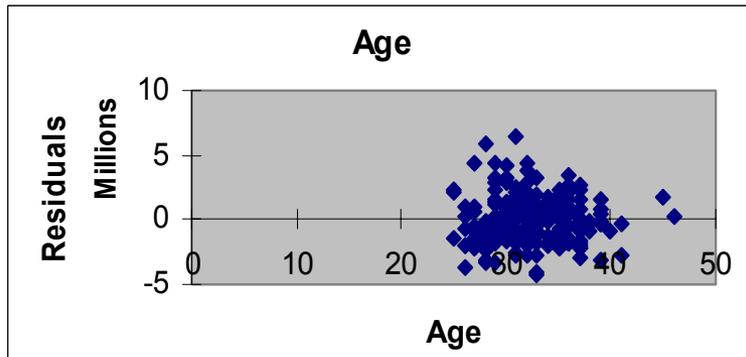
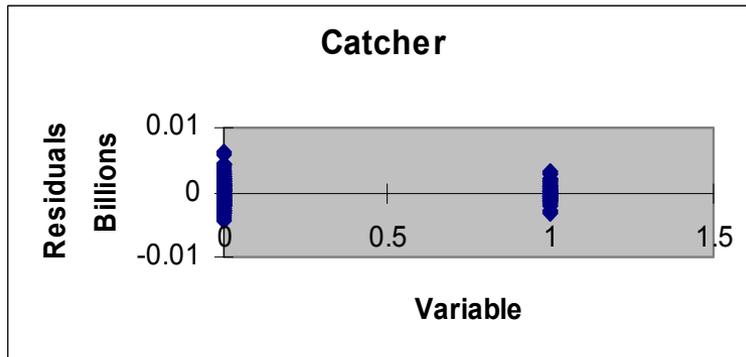
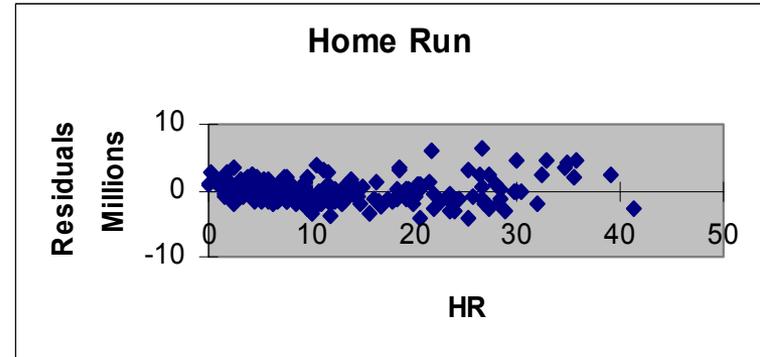
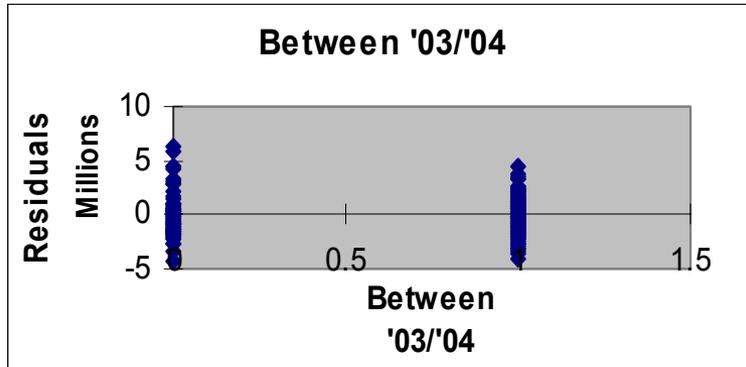
Predicted Salary vs. Actual Salary



Residuals Analysis



Residuals Analysis



Factors Outside our Project's Scope

- Performance in-the-clutch:
 - Batting average with runners in scoring position
 - Newer statistics: “Close & Late”, etc.
- Injury History:
 - # days spent on disabled list
- Leadership:
 - Make-up
 - Personality factors that contribute to likely future performance (lifestyle, etc.)
 - Impact on others - “Clubhouse guy” effect
 - “Leaders”: Miguel Tejada, David Ortiz
 - “Burdens”: A.J. Pierzynski, Barry Bonds
- Other:
 - Native country and language
 - Agent (a.k.a. “The Scott Boras factor”)
 - Scarcity of particular positions in free agent market will likely make a difference.

Additional improvements

- Easier access to data would allow
 - Trying different performance variables
 - Trying new physical/team variables
 - More years of data
- We could select better stats
 - Defense
 - No good statistics publicly available
 - FPCT better captures errors relative to total chances
 - Still skewed by range factors
 - Zone Rating, Gold Glove (fielding excellence) awards
 - Performance
 - MVP awards, Silver Slugger awards
- We could expand the coverage of the model
 - Minor League Players
 - Amateur Players
 - Pitchers?
- Upon (last minute) reflection, it appears that a model swapping out some variables with exponential variables might improve the model. How do you determine what power to use?

Conclusions

- OBP or walks do not necessarily matter much.
- Teams pay the market rate to their own players (no premium or discount).
 - Although some players never make it to free agency
- Catchers get paid \$1.6M more on average than other batting positions.
- Contract length was very highly correlated to player salary.
- 71.3% of the variability in FA player salary can be predicted by the simplified model:

$$Y = -6.53 - 1.29*(\text{'03/'04}) + 1.55*\text{Catcher} - 0.16*\text{Age} \\ + 0.22*\text{HRs} + 0.10*\text{SBs} + 43.48*\text{Average}$$

Who are we?



Top 5 Overpaid and Underpaid Players

On a \$ Basis					
Overpaid			Y^	Y	Difference
1	Magglio	Ordonez	8,649,203	15,000,000	-6,350,797
2	Troy	Glaus	5,377,806	11,250,000	-5,872,194
3	Miguel	Tejada	7,583,804	12,000,000	-4,416,196
4	Carlos	Beltran	12,590,581	17,000,000	-4,409,419
5	Carlos	Delgado	8,631,951	13,000,000	-4,368,049
Underpaid					
1	Raul	Mondesi	5,293,958	1,000,000	4,293,958
2	Raul	Mondesi	5,793,857	1,750,000	4,043,857
3	Luis	Matos	4,636,856	975,000	3,661,856
4	Brad	Fullmer	4,341,003	1,000,000	3,341,003
5	A.J.	Pierzynski	5,531,998	2,250,000	3,281,998

On a % Basis					
Overpaid			Y^	Y	Difference
1	Mike	Mordecai	-2,160,114	500,000	-532%
2	Enrique	Wilson	-2,593,737	700,000	-471%
3	Brian	Banks	-1,456,898	400,000	-464%
4	Frank	Menechino	-1,437,743	400,000	-459%
5	Joe	McEwing	-1,175,150	500,000	-335%
Underpaid					
1	Raul	Mondesi	5,293,958	1,000,000	429%
2	Ellis	Burks	3,855,035	750,000	414%
3	Luis	Matos	4,636,856	975,000	376%
4	Todd	Greene	3,422,980	750,000	356%
5	Ramon	Castro	1,776,307	400,000	344%

What would Juran Make?

- Suppose Juran made it to the big leagues...
 - Signed in '05/'06 (gets one more year of inflation)
 - Plays catcher
 - Hits a couple of home runs
 - Gets picked off regularly
 - Just misses the Mendoza line

Between 03/'04	Catcher	Age	Home Runs	Stolen Bases	Average
-1	1	63	2	-5	0.199

Juran's Final Salary	-\$5,088,861
-----------------------------	---------------------

- It's no surprise that Juran would in fact have to pay MLB to play!