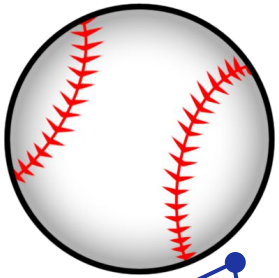
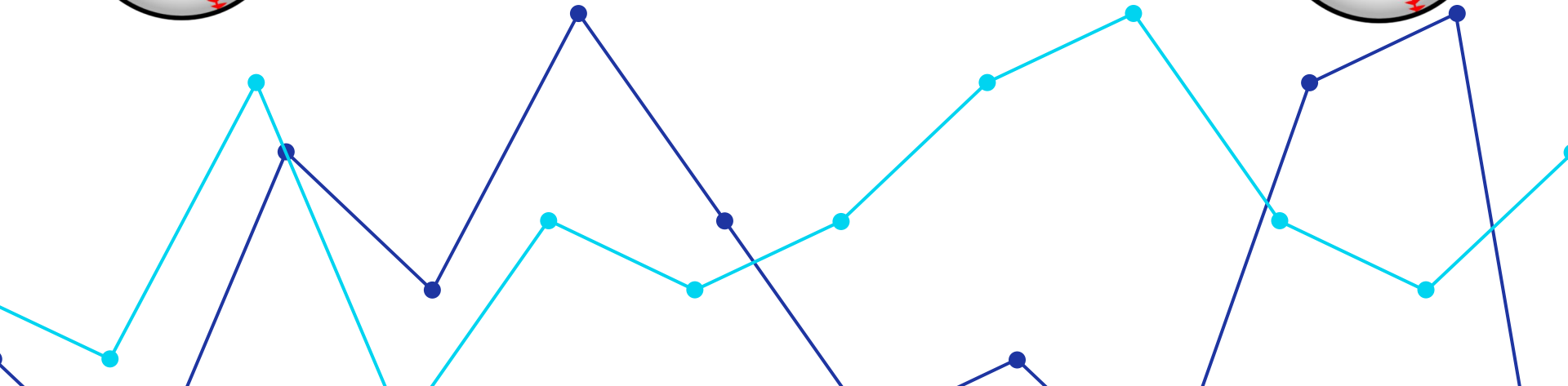
How Do Game Statistics Affect Player Salary?

DSC 205 Final Project Presentation



Rachel Matherly & Princess Allotey



Have you played baseball/softball before?

Background

- There is an assumption that baseball players with good statistics are paid more

- Is this assumption really true?

Dataset

2010MLBSalary

OpenIntro Website

MLBSTATS2010

Rotowire Website

player	team	position	salary	Team	Pos	Age	G	AB	R	H	X2B	X3B	HR	RBI	SB	CS	BB	SO	SH	SF	HBP	AVG	OBP
Aaron Hill	Toronto Blue Jays	Second Baseman	4000.000	TOR	2B	38	138	528	70	108	22	0	26	68	2	2	41	85	1	2	8	0.205	0.271
Aaron Miles	Cincinnati Reds	Second Baseman	2700.000	STL	2B	43	79	139	14	39	5	0	0	9	0	1	6	14	3	2	1	0.281	0.311
Aaron Rowand	San Francisco Giants	Outfielder	13600.000	SF	OF	43	105	331	42	76	12	2	11	34	5	3	16	74	1	1	8	0.230	0.281
Adam Dunn	Washington Nationals	First Baseman	12000.000	WAS	1B	41	158	558	85	145	36	2	38	103	0	1	77	199	0	4	9	0.260	0.356
Adam Everett	Detroit Tigers	Shortstop	1550.000	DET	SS	43	31	81	6	15	5	0	0	4	2	1	4	18	3	1	0	0.185	0.221
Adam Jones	Baltimore Orioles	Outfielder	465.000	BAL	OF	35	149	581	76	165	25	5	19	69	7	7	23	119	2	2	13	0.284	0.325
Adam Kennedy	Washington Nationals	Second Baseman	1250.000	WAS	3B	44	135	342	43	85	16	1	3	31	14	2	37	44	1	4	5	0.249	0.327
Adam LaRoche	Arizona Diamondbacks	First Baseman	4500.000	ARI	1B	41	151	560	75	146	37	2	25	100	0	1	48	172	0	4	3	0.261	0.320
Adam Lind	Toronto Blue Jays	Designated Hitter	550.000	TOR	1B	37	150	569	57	135	32	3	23	72	0	0	38	144	0	3	3	0.237	0.287
Adam Moore	Seattle Mariners	Catcher	401.000	SEA	C	36	60	205	12	40	6	0	4	15	0	1	8	63	1	2	2	0.195	0.230
Adam Rosales	Oakland Athletics	Third Baseman	410.000	OAK	2B	37	80	255	31	69	8	2	7	31	2	2	19	65	2	2	1	0.271	0.321
Adrian Beltre	Boston Red Sox	Third Baseman	9000.000	BOS	3B	41	154	589	84	189	49	2	28	102	2	1	40	82	0	7	5	0.321	0.365
Adrian Gonzalez	San Diego Padres	First Baseman	4875.000	SD	1B	38	160	591	87	176	33	0	31	101	0	0	93	114	2	4	2	0.298	0.393
A.J. Pierzynski	Chicago White Sox	Catcher	6750.000	CWS	C	43	128	474	43	128	29	0	9	56	3	4	15	39	6	2	6	0.270	0.300
Akinori Iwamura	Pittsburgh Pirates	Second Baseman	4850.000	OAK	2B	41	10	31	3	4	1	0	0	4	0	0	5	10	0	0	0	0.129	0.250

Questions

Question 1

Can we create a model to determine the salary of baseball players based on their game statistics?

Question 2

How effective are each of our predictor variables at determining salary?

Question 3

What predictor variables are most significant in determining salary?

Question 4

How could we change our model to decrease error?

Abbreviations

- Salary: yearly salary in thousands of dollars
- Team: out of 30 Major League Baseball (MLB) teams
- Pos: field position (C,P, 1B, 2B, 3B, SS, OF, DH)
- Age
- G: number of games played
- AB: at-bats
- R: runs
- H: hits
- X2B: doubles
- X3B: triples
- HR: homeruns
- BB: Walks
- RBI: runs-batted-in
- SB: stolen bases
- CS: caught stealing
- SH: sacrifice hits
- SF: sacrifice flyouts
- HBP: hit-by-pitch
- AVG: batting average
- OBP: on-base percentage
- SLG: slugging percentage
- OPS: the sum of on-base percentage and slugging average
- SO: Strikeouts

First-Order Linear Model, Model Utility F-Test, and βi t-tests

```
> my_lm=with(fulldata, lm(salary~.,fulldata[,5:26]))
> summary(my_lm)

Call:
lm(formula = salary ~ ., data = fulldata[, 5:26])

Residuals:
   Min     1Q Median     3Q    Max
 -9174  -2192   -286   1616  15502

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15136.645   2727.096  -5.550 5.38e-08 ***
TeamATL        753.837   1423.563   0.530  0.59674
TeamBAL        225.030   1469.375   0.153  0.87836
TeamBOS       1553.416   1387.113   1.120  0.26347
TeamCHC       2610.081   1404.617   1.858  0.06392 .
TeamCIN      -1806.059   1428.819  -1.264  0.20700
TeamCLE      -1000.027   1400.692  -0.714  0.47570
TeamCOL        -68.538   1485.550  -0.046  0.96323
TeamCWS        -71.027   1469.526  -0.048  0.96148
TeamDET       2614.699   1416.387   1.846  0.06567 .
TeamFLA         17.181   1432.135   0.012  0.99043
TeamHOU       1029.669   1481.276   0.695  0.48741
TeamKC       -1010.529   1479.748  -0.683  0.49509
TeamLAA         83.871   1456.429   0.058  0.95411
TeamLAD       -264.641   1359.143  -0.195  0.84572
TeamMIL      -1328.916   1420.253  -0.936  0.35003
TeamMIN        850.844   1454.884   0.585  0.55902
TeamNYM       1151.280   1450.082   0.794  0.42773
TeamNYY       4250.675   1446.848   2.938  0.00351 **
TeamOAK        445.957   1434.754   0.311  0.75611
TeamPHI       1105.739   1458.022   0.758  0.44870
TeamPIT       -107.959   1403.412  -0.077  0.93872
TeamSD       -1199.042   1435.029  -0.836  0.40394
TeamSEA        927.188   1497.318   0.619  0.53614
TeamSF        1175.660   1348.396   0.872  0.38382
TeamSTL        528.474   1403.040   0.377  0.70664
TeamTB         966.969   1432.680   0.675  0.50013
TeamTEX        650.940   1418.386   0.459  0.64655
TeamTEX        650.940   1418.386   0.459  0.64655
TeamTOR       -735.618   1490.563  -0.494  0.62193
TeamWAS        834.315   1453.617   0.574  0.56634
Pos2B         -783.693    811.879  -0.965  0.33502
Pos3B         -536.150    782.346  -0.685  0.49357
PosC         -1830.187    772.644  -2.369  0.01835 *
PosDH         1653.412   1072.114   1.542  0.12386
PosOF           44.714    673.950   0.066  0.94714
PosSS          731.660    891.119   0.821  0.41213
Age            364.964     46.959   7.772 7.42e-14 ***
G              -79.723     13.814  -5.771 1.64e-08 ***
AB              38.979      9.274   4.203 3.29e-05 ***
R              -20.402     33.896  -0.602  0.54760
H              -75.766     32.287  -2.347  0.01946 *
X2B             44.168     51.015   0.866  0.38716
X3B            -43.175    132.961  -0.325  0.74557
HR             114.780     87.810   1.307  0.19196
RBI             36.224     32.443   1.117  0.26490
SB             -49.906     40.422  -1.235  0.21774
CS             258.847    117.101   2.210  0.02767 *
BB              55.648     22.603   2.462  0.01426 *
SO             -28.769     11.714  -2.456  0.01450 *
SH            -181.477     98.674  -1.839  0.06668 .
SF             -44.316    123.731  -0.358  0.72042
HBP             -9.354     68.784  -0.136  0.89190
AVG          16393.789  11418.563   1.436  0.15191
OBP           4567.500   9579.616   0.477  0.63379
SLG          -5137.747   4860.321  -1.057  0.29115
OPS                 NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3654 on 377 degrees of freedom
Multiple R-squared:  0.4996,    Adjusted R-squared:  0.4279
F-statistic: 6.969 on 54 and 377 DF,  p-value: < 2.2e-16
```

Stepwise Linear Regression

- **Forward**
  - `lm(formula = salary ~ RBI + Age + G + BB + AB + Pos + SO + SH + H + AVG, data = fulldata)`
- **Backward**
  - `lm(formula = salary ~ Pos + Age + G + AB + H + RBI + SB + CS + BB + SO + SH + AVG, data = fulldata)`
- **Both**
  - `lm(formula = salary ~ RBI + Age + G + BB + AB + Pos + SO + SH + H + AVG, data = fulldata)`

Multicollinearity

```
> with(fulldata, cor(data.frame(Age, G , AB ,H , RBI , BB , SO , SH , AVG)))
             Age          G          AB          H         RBI          BB          SO          SH          AVG
Age  1.00000000 -0.1257821 -0.1371007 -0.1268507 -0.08311389 -0.09046987 -0.19251198 -0.13177165 -0.02880056
G   -0.12578207  1.0000000  0.9493950  0.9230564  0.84693711  0.79556327  0.82592709  0.22546066  0.48820995
AB  -0.13710070  0.9493950  1.0000000  0.9846682  0.88674899  0.81905268  0.83619544  0.22139227  0.51252822
H   -0.12685070  0.9230564  0.9846682  1.0000000  0.89370158  0.80806289  0.79628110  0.19510048  0.59229424
RBI -0.08311389  0.8469371  0.8867490  0.8937016  1.00000000  0.81555431  0.81674062 -0.06531576  0.50816242
BB  -0.09046987  0.7955633  0.8190527  0.8080629  0.81555431  1.00000000  0.80001078  0.08212040  0.42374570
SO  -0.19251198  0.8259271  0.8361954  0.7962811  0.81674062  0.80001078  1.00000000  0.05919779  0.35452218
SH  -0.13177165  0.2254607  0.2213923  0.1951005 -0.06531576  0.08212040  0.05919779  1.00000000  0.05602567
AVG -0.02880056  0.4882100  0.5125282  0.5922942  0.50816242  0.42374570  0.35452218  0.05602567  1.00000000
```

Nested F-Test

```
> anova(steplm, newlm)
Analysis of Variance Table

Model 1: salary ~ RBI + Age + G + BB + AB + Pos + SO + SH + H + AVG
Model 2: salary ~ Pos + Age + AB + H + BB + SO + SH + AVG
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1    416 5727883813
2    418 6346964172 -2 -619080359 22.481 5.359e-10 ***
```

Stepwise Linear Regression Part 2

```
Step:  AIC=7154.36
salary ~ BB + Age + SH + H + SO + Pos

       Df  Sum of Sq        RSS    AIC
<none>                6363818832 7154.4
+ AB    1    8777788 6355041045 7155.8
- Pos   6  202673020 6566491852 7155.9
+ AVG   1    1482449 6362336383 7156.3
- SO    1   68357928 6432176760 7157.0
- H     1  183921020 6547739852 7164.7
- BB    1  278171918 6641990751 7170.8
- SH    1  282968509 6646787341 7171.2
- Age   1 1026133060 7389951892 7216.9

Call:
lm(formula = salary ~ BB + Age + SH + H + SO + Pos, data = fulldata)
```

*This took out AB and AVG

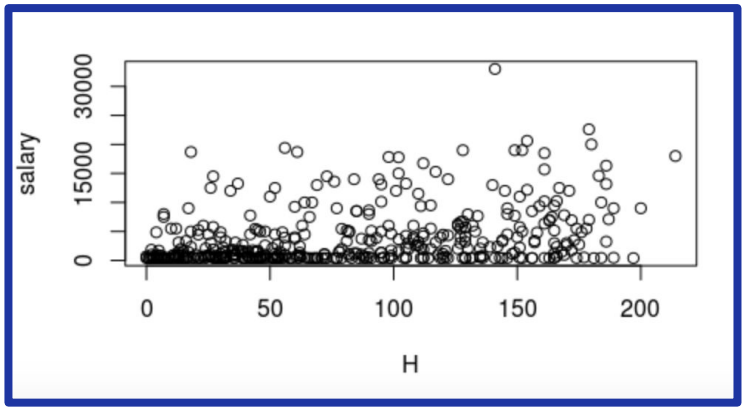
T-tests with Interactions and another Nested F-test

```
Age:SH       -53.2781    20.2935  -2.625  0.00900 **
Age:H          5.8996     1.8742   3.148  0.00177 **
```

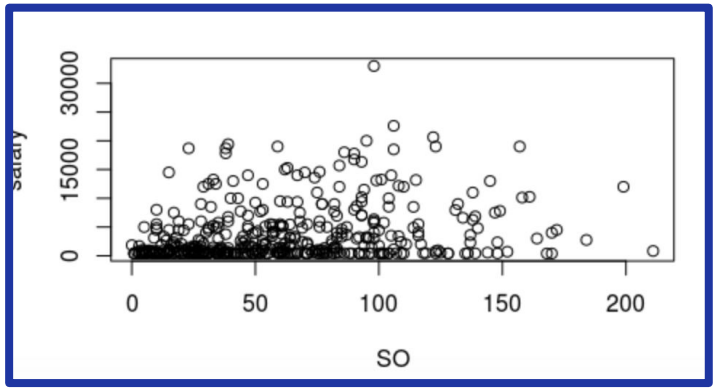
```
> anova(simplelm, improvelm)
Analysis of Variance Table

Model 1: salary ~ BB + Age + SH + H + SO + Pos
Model 2: salary ~ BB + Age + SH + H + SO + Pos + (Age * SH) + (Age * H)
  Res.Df        RSS Df Sum of Sq     F   Pr(>F)
1    420 6363818832
2    418 5694638798  2 669180034 24.56 8.23e-11 ***
```

Scatterplots

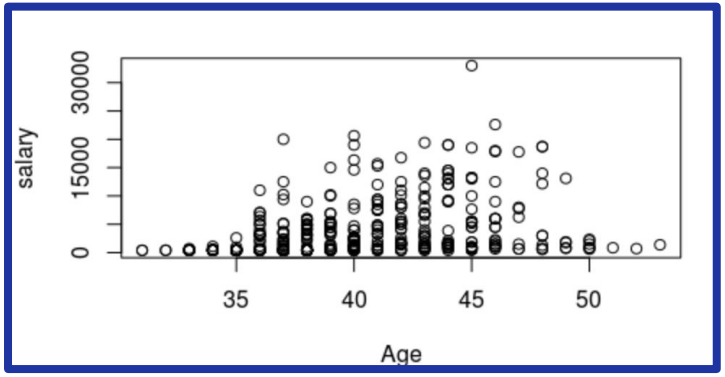
Hits vs. Salary



SO vs. Salary



Age vs. Salary



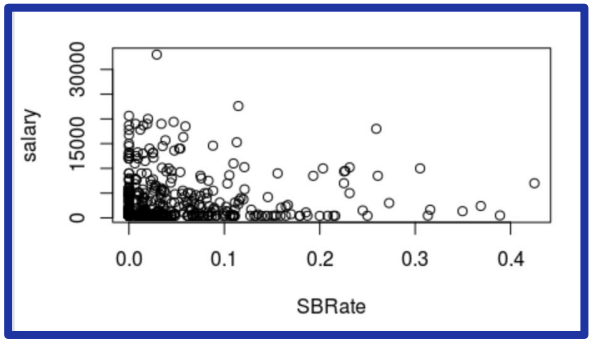
F-Test for Model Utility for Rate Linear Model and Stepwise Regression

```
TeamNYM      1.275e+03  1.449e+03   0.879 0.379741
TeamNYY      3.833e+03  1.428e+03   2.684 0.007592 **
TeamOAK      2.756e+02  1.438e+03   0.192 0.848065
TeamPHI      9.826e+02  1.463e+03   0.672 0.502140
TeamPIT      3.388e+02  1.414e+03   0.240 0.810740
TeamSD      -1.398e+03  1.436e+03  -0.973 0.331086
TeamSEA      9.855e+02  1.494e+03   0.660 0.509954
TeamSF       1.659e+03  1.357e+03   1.222 0.222395
TeamSTL      6.693e+02  1.394e+03   0.480 0.631413
TeamTB       4.300e+02  1.426e+03   0.302 0.763179
TeamTEX      6.742e+02  1.405e+03   0.480 0.631560
TeamTOR      1.760e+02  1.484e+03   0.119 0.905639
TeamWAS      2.512e+02  1.458e+03   0.172 0.863302
Pos2B       -1.137e+03  8.169e+02  -1.392 0.164770
Pos3B       -7.284e+02  7.734e+02  -0.942 0.346892
PosC        -2.234e+03  7.705e+02  -2.899 0.003957 **
PosDH        1.206e+03  1.082e+03   1.115 0.265375
PosOF       -2.452e+02  6.703e+02  -0.366 0.714656
PosSS        2.388e+02  8.884e+02   0.269 0.788241
Age          3.728e+02  4.766e+01   7.821 5.15e-14 ***
G           -5.692e-01  5.542e+00  -0.103 0.918241
AVG          5.247e+04  2.572e+04   2.040 0.041997 *
OBP         -5.936e+04  2.206e+04  -2.690 0.007454 **
SLG         -3.935e+03  4.737e+03  -0.831 0.406688
OPS                 NA         NA      NA       NA
HRate        3.216e+03  4.742e+03   0.678 0.498050
ABRate      -8.261e+02  1.227e+03  -0.673 0.501272
RRate        4.048e+02  2.839e+03   0.143 0.886706
RBIRate      6.675e+03  2.216e+03   3.012 0.002764 **
SBRate      -4.502e+03  3.601e+03  -1.250 0.212033
BBRate       1.732e+04  4.743e+03   3.651 0.000297 ***
SORate      -3.960e+02  1.085e+03  -0.365 0.715474
HBPRate      1.517e+04  7.667e+03   1.979 0.048582 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3712 on 383 degrees of freedom
Multiple R-squared:  0.4752,    Adjusted R-squared:  0.4094
F-statistic: 7.225 on 48 and 383 DF,  p-value: < 2.2e-16
```

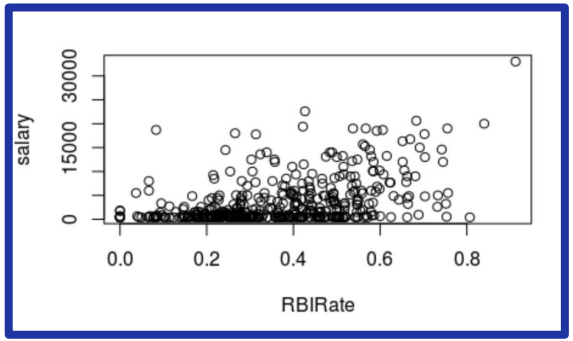
```
Step:  AIC=7128.45
salary ~ RBIRate + Age + BBRate + Pos + HRate + OPS + SBRate

          Df  Sum of Sq        RSS    AIC
<none>                  5965696517 7128.5
- SBRate   1   37797217 6003493734 7129.2
+ RRate    1   17131723 5948564795 7129.2
+ ABRate   1   13158778 5952537740 7129.5
+ AVG      1    8408086 5957288431 7129.8
+ SORate   1    8327146 5957369372 7129.9
- OPS      1   50547581 6016244098 7130.1
+ OBP      1    2978606 5962717911 7130.2
+ SLG      1    2978606 5962717911 7130.2
+ HBPRate  1    1990991 5963705526 7130.3
+ G        1      88326 5965608191 7130.4
- HRate    1  157804742 6123501259 7137.7
- Pos      6  343715007 6309411524 7140.7
+ Team    29  587821721 5377874796 7141.6
- RBIRate  1  234282950 6199979467 7143.1
- BBRate   1  270699670 6236396187 7145.6
- Age      1 1083884098 7049580615 7198.6

Call:
lm(formula = salary ~ RBIRate + Age + BBRate + Pos + HRate +
    OPS + SBRate, data = fulldata)
```

Correlation and More Scatterplots!

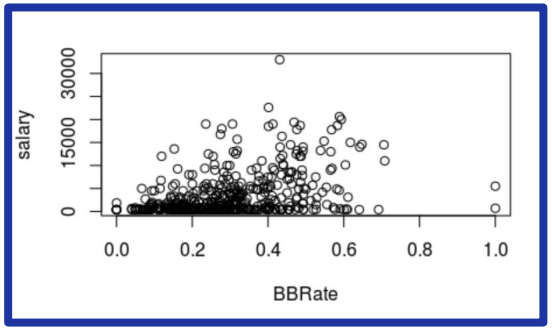
SB Rate vs. Salary



RBI Rate vs. Salary



BB Rate vs. Salary




```
> with(fulldata, cor(data.frame(RBIRate , Age , BBRate , HRate , OPS , SBRate)))
            RBIRate          Age     BBRate      HRate       OPS      SBRate
RBIRate  1.00000000 -0.021486010 0.48648779  0.70491137 0.710032967  0.02172648
Age     -0.02148601  1.000000000 0.02050708 -0.06795965 0.002340383 -0.16229277
BBRate   0.48648779  0.020507081 1.00000000  0.41193907 0.456122337  0.14308288
HRate    0.70491137 -0.067959654 0.41193907  1.00000000 0.644634701  0.34682788
OPS      0.71003297  0.002340383 0.45612234  0.64463470 1.000000000  0.07716187
SBRate   0.02172648 -0.162292766 0.14308288  0.34682788 0.077161865  1.00000000
```

T-tests for βi and Another Nested F-test

```
> newratelm=with(fulldata, lm(salary~RBIRate + Age + BBRate + Pos + HRate + OPS + SBRate))
> summary(newratelm)

Call:
lm(formula = salary ~ RBIRate + Age + BBRate + Pos + HRate +
    OPS + SBRate)

Residuals:
    Min      1Q  Median      3Q     Max
-9719.7 -2210.5  -396.9  1628.0 19949.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -15968.41    2165.37  -7.374 8.91e-13 ***
RBIRate       7514.67    1852.52   4.056 5.94e-05 ***
Age            387.45      44.41   8.725  < 2e-16 ***
BBRate        6228.33    1428.40   4.360 1.64e-05 ***
Pos2B         -871.21     781.87  -1.114 0.265803
Pos3B         -973.34     761.34  -1.278 0.201798
PosC         -2345.07     736.40  -3.184 0.001558 **
PosDH         1527.68    1049.32   1.456 0.146173
PosOF         -132.47     656.16  -0.202 0.840105
PosSS          338.64     864.77   0.392 0.695559
HRate         3890.42    1168.58   3.329 0.000948 ***
OPS          -3933.72    2087.74  -1.884 0.060230 .
SBRate       -5316.93    3263.28  -1.629 0.103997
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3773 on 419 degrees of freedom
Multiple R-squared:  0.4068,    Adjusted R-squared:  0.3898
F-statistic: 23.94 on 12 and 419 DF,  p-value: < 2.2e-16
```

```
> anova(newratelm, noposlm)
Analysis of Variance Table

Model 1: salary ~ RBIRate + Age + BBRate + Pos + HRate + OPS + SBRate
Model 2: salary ~ RBIRate + Age + BBRate + HRate + OPS + SBRate
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1    419 5965696517
2    425 6309411524 -6 -343715007 4.0235 0.0006232 ***
```

Stepwise Linear Regression and Nested F-test...Again

```
Step:  AIC=7139.64
salary ~ RBIRate + Age + BBRate + HRate + OPS

          Df  Sum of Sq        RSS    AIC
<none>                  6323783061 7139.6
- OPS      1   35171347 6358954408 7140.0
+ SBRate   1   14371537 6309411524 7140.7
- HRate    1  141631749 6465414810 7147.2
- RBIRate  1  315633892 6639416953 7158.7
- BBRate   1  336159839 6659942900 7160.0
- Age      1 1237750414 7561533475 7214.9

Call:
lm(formula = salary ~ RBIRate + Age + BBRate + HRate + OPS, data = fulldata)
```

*This took out SB rate

```
> anova(noposlm, improvelm2)
Analysis of Variance Table

Model 1: salary ~ RBIRate + Age + BBRate + HRate + OPS + SBRate
Model 2: salary ~ RBIRate + Age + BBRate + HRate + OPS
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    425 6309411524
2    426 6323783061 -1 -14371537 0.9681 0.3257
```

Prediction Intervals

Non-Rate linear model

Player Name	Lower Prediction Interval	Upper Prediction Interval	Actual
Derek Jeter	7206.272	22211.04	22600

Rate linear model

Player Name	Lower Prediction Interval	Upper Prediction Interval	Actual
Derek Jeter	642.4682	15909.68	22600

Prediction Intervals

Non-Rate linear model

Player Name	Lower Prediction Interval	Upper Prediction Interval	Actual
Drew Stubbs	-5598.187	9159.288	400

Rate linear model

Player Name	Lower Prediction Interval	Upper Prediction Interval	Actual
Drew Stubbs	-4400.862	10907.52	400

Prediction Intervals

Non-Rate linear model

Player Name	Lower Prediction Interval	Upper Prediction Interval	Actual
Hunter Pence	-3703.747	10987.25	3500

Rate linear model

Player Name	Lower Prediction Interval	Upper Prediction Interval	Actual
Hunter Pence	-3079.254	12154.83	3500

Prediction Intervals

Non-Rate linear model

Player Name	Lower Prediction Interval	Upper Prediction Interval	Actual
Ichiro Suzuki	6513.436	21608.21	18000

Rate linear model

Player Name	Lower Prediction Interval	Upper Prediction Interval	Actual
Ichiro Suzuki	-1453.367	14035.7	18000

Conclusions

Non-Rate Linear Model:

$$y=385.78+62.82x_1+37.79x_2+1491.13x_3-206.2x_4-16.19x_5-760x_6-1079.53x_7-1514.69x_8+1784.81x_9-442.37x_{10}+192.15x_{11}-46.76x_{12}+5.79x_{13}$$

y=salary
x_1=BB
x_2=Age
x_3=SH
x_4=H
x_5=SO
x_6=Pos2B
x_7=Pos3B
x_8=PosC
x_9=PosDH
x_{10}=PosOF
x_{11}=PosSS
x_{12}=Age*SH
x_{13}=Age*H

	R^2	R_a^2
Non-Rate Model	0.4337	0.4161
Rate Model	0.3726	0.3637

Limitations

- Salary dataset focused on 672 players from 2010 which could be estimates from contracts made in a previous year
- Final dataset had 432 total entries which represented only batters
- Our model might provide some indication of salaries in future years, but financial circumstances of the league and individual teams change from year to year

References

- https://www.nytimes.com/interactive/2014/02/13/sports/baseball/jeter-long-lived-greatness.html
- https://www.openintro.org/data/index.php?data=mlb
- https://www.rotowire.com/baseball/stats.php?season=2010
- https://www.si.com/mlb/2020/07/22/richest-mlb-contracts-mookie-betts-bryce-harper-mike-trout
- https://thetrovesportsden.com/products/drew-stubbs-items
- https://www.stickpng.com/img/sports/baseball/major-league-baseball-mlb/san-francisco-giants/san-francisco-giants-hunter-pence
- https://grandsalami.net/posts/ichiro-forever/

Any Questions?