A Predictive Model:
The Effect of Major League Baseball Player Statistics on their Future Salary

Raaid Ahmad
Johns Hopkins University
Applied Statistics and Data Analysis
April 3, 2006

Table of Contents

I. Introduction

A. The Data

The data consists of 122 observations of player statistics and salary information for Major League Baseball free agents at the end of the 2004 and 2005 season. Additionally, the salary for any given observation is the salary given to the player *after* they had earned the statistics in the observation. For example, if the statistics in a given observation are from the year 2004, the salary is how much the player got paid in the year 2005.

AB	R	H.1B	H.2B	H.3B	HR	RBI	SB	CS	BB	BA	OBP	SLG	OPS	SALARY	P.1B	P.2B	P.3B	P.SS	P.C
379	68	117	17	3	7	38	7	2	31	0.309	0.366	0.425	0.791	$2,050,000	0	0	1	0	0
464	52	119	19	0	7	59	0	1	51	0.256	0.334	0.343	0.677	$750,000	1	0	0	0	0
369	69	100	19	3	4	21	36	10	51	0.271	0.371	0.371	0.743	$1,500,000	0	1	0	0	0
374	49	117	24	3	12	53	1	0	17	0.313	0.348	0.489	0.837	$3,250,000	0	0	0	0	0
616	100	175	31	11	12	58	46	10	62	0.284	0.348	0.429	0.777	$13,000,000	0	0	0	1	0
233	30	64	12	1	9	42	4	0	27	0.275	0.348	0.451	0.799	$1,100,000	1	0	0	0	0
412	49	93	24	3	10	40	7	2	32	0.226	0.294	0.371	0.665	$2,250,000	0	0	0	0	0
624	117	197	35	6	10	75	18	1	53	0.316	0.366	0.439	0.805	$13,000,000	0	0	0	0	0
519	69	153	34	3	10	62	0	0	59	0.295	0.369	0.43	0.799	$4,750,000	0	0	1	0	0
449	57	122	28	1	9	50	0	1	54	0.272	0.355	0.399	0.753	$2,100,000	1	0	0	0	0

Note: From this point, the salary will be in thousands of dollars and the percentage values (BA, OBP, SLG, OPS) will be multiplied by 100 to reflect percentages and maintain the integrity of polynomial terms.

Finally, since there is a legitimate possibility that some players get paid more based on their position, some dummy variables were added to represent a player's position.

P.1B = 1 if the player is a first baseman, 0 otherwise
P.2B = 1 if the player is a second baseman, 0 otherwise
P.3B = 1 if the player is a third baseman, 0 otherwise
P.SS = 1 if the player is a shortstop, 0 otherwise
P.C = 1 if the player is a catcher, 0 otherwise
P.1B = P.2B = P.3B = P.SS = P.C = 0 implies that the player is an outfielder.

In order to prevent a bias as much as possible by selecting certain players, (more popular players, players on certain teams, etc) a list of *all* free agents after 2004 and 2005 was created. From this list, all the hitters were extracted and then ESPN.com data was collected for each player's statistics for the year at the end of which he would become a free agent. After that, the salaries for the exact same players (for the following year) were found and "joined" with the corresponding player statistics. Finally, to check for a bias between salaries of different years, an average was taken and luckily, the average salaries for free agents for both 2004 and 2005 was *very* similar (a difference of 0.3%).

B. The Objective(s)

The dependent variable that will be used is SALARY. The major goal here is to determine whether statistics from a player's most recent year has a major affect on their salary for the upcoming year. If this is the case, MLB teams may use this as a model to determine how much a player's true value is. This can also let MLB teams know if current players' contracts need to be re-worked or if they are being paid their true value. In essence, monetary value can be assigned to each unit of a baseball statistic. If the player's most recent statistics do not truly have an effect on their upcoming year's salary then it would be important to know this, and to find out what factors *do* affect salary.

Even more specifically, this model can allow teams, managers, and players to see what types of skills are rewarded. The comparison between compensation for hitting for power, hitting for average, speed, and defense can be made with this model.

C. Summary

There will be 15 independent variables:

- AB – at bats
- R – runs scored
- H.1B – hits
- H.2B – doubles
- H.3B – triples
- HR – home runs
- RBI – runs batted in
- SB – stolen bases
- CS – caught stealing
- BB – walks
- BA – batting average
- OBP – on-base percentage
- SLG – slugging percentage
- OPS – on-base plus slugging percentage
- Position (represented by dummy variables P.1B, P.2B, P.3B, P.SS, P.C)

These will be used in an attempt to predict the SALARY, the independent variable, of a player for the following year.

II. Exploratory Data Analysis

A. Multicollinearity

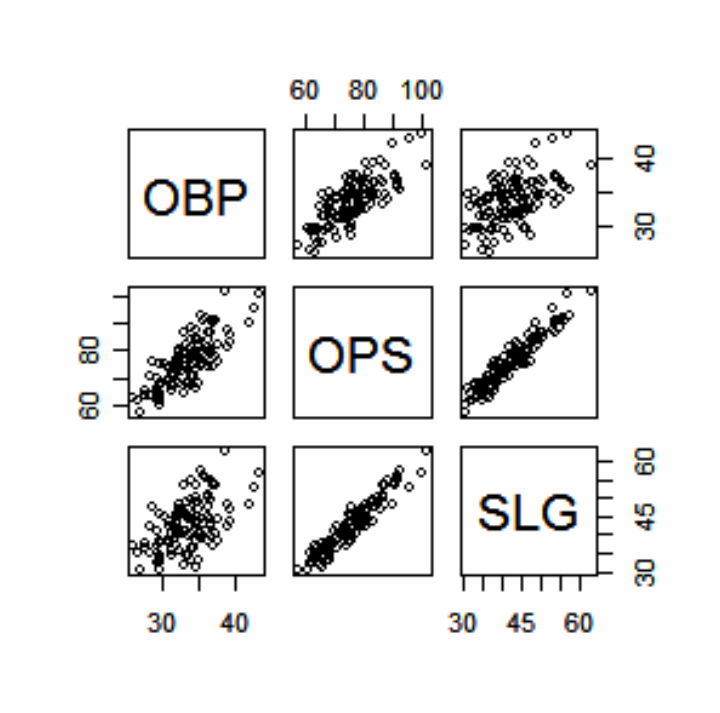
In order to possibly eliminate some variables that have high levels of correlation, the correlation matrix and the scatterplot matrices for the data can be used. For the purposes of removing some subjectivity from these measurements, a threshold of .85 was selected as denoting sufficient multicollinearity for variable removal. In other words, any variables with correlation coefficient > .85 or < -.85 will be considered highly correlated.

	AB	R	H.1B	H.2B	H.3B	HR	RBI	BB
AB	1.0000000	0.8792424	0.9678020	0.8302682	0.3504687	0.5377756	0.7386989	0.6036225
R	0.8792424	1.0000000	0.8959399	0.7515438	0.4844153	0.6357878	0.7313034	0.7289972
H.1B	0.9678020	0.8959399	1.0000000	0.8273880	0.3734507	0.5193076	0.7377099	0.5930149
H.2B	0.8302682	0.7515438	0.8273880	1.0000000	0.2692461	0.4613994	0.6661123	0.5397785
H.3B	0.3504687	0.4844153	0.3734507	0.2692461	1.0000000	0.1201148	0.2394769	0.3977688
HR	0.5377756	0.6357878	0.5193076	0.4613994	0.1201148	1.0000000	0.8672076	0.5463511
RBI	0.7386989	0.7313034	0.7377099	0.6661123	0.2394769	0.8672076	1.0000000	0.5870659
BB	0.6036225	0.7289972	0.5930149	0.5397785	0.3977688	0.5463511	0.5870659	1.0000000

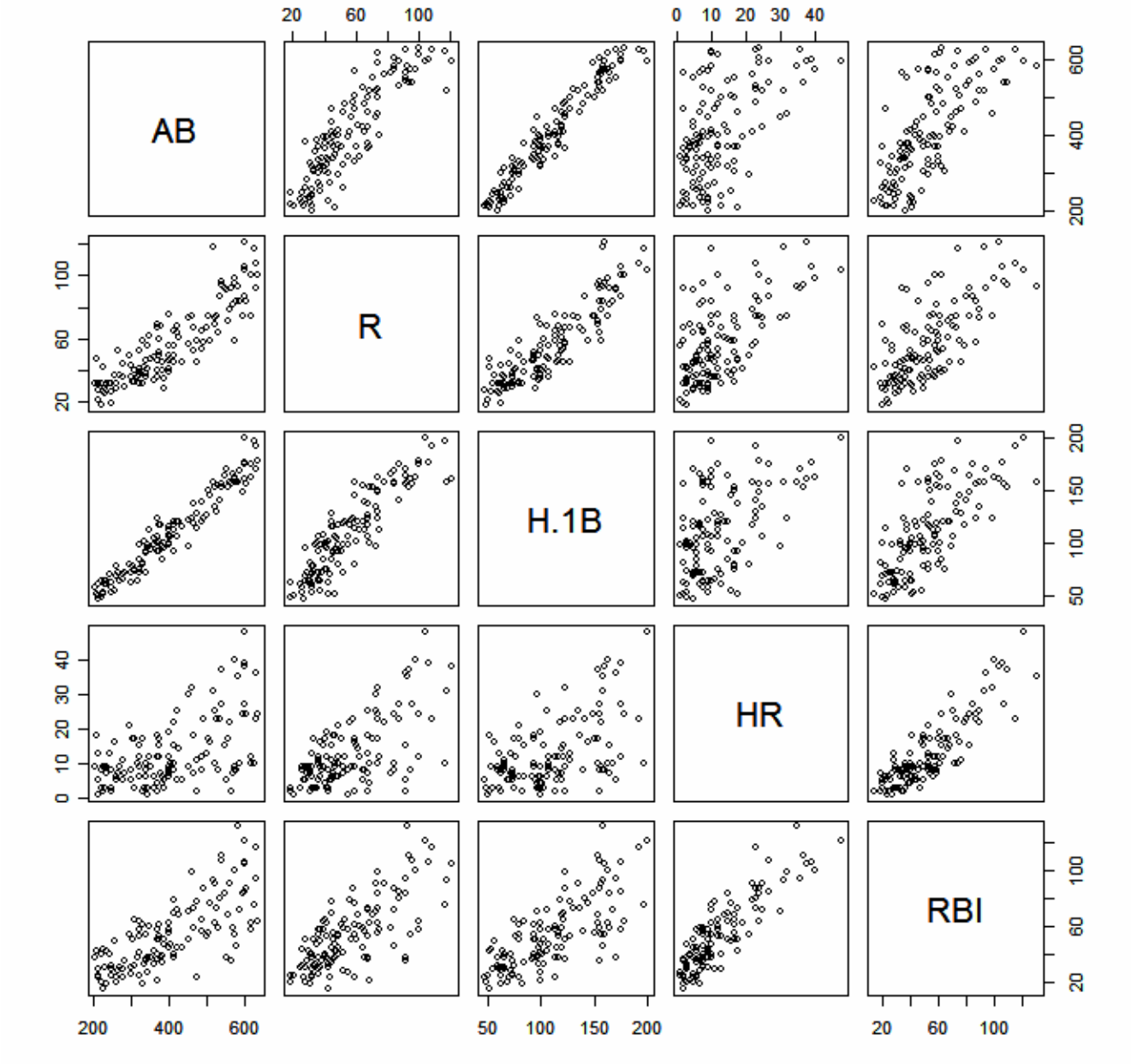
	H.1B	BA	OBP	SLG	OPS	SB	CS
H.1B	1.0000000	0.5540542	0.37729702	0.357401022	0.40814017	0.381125584	0.32491821
BA	0.5540542	1.0000000	0.74165674	0.479626201	0.63446988	0.169071833	0.01683280
OBP	0.3772970	0.7416567	1.00000000	0.538309206	0.77465868	0.239361423	0.09642614
SLG	0.3574010	0.4796262	0.53830921	1.000000000	0.94992665	-0.002978809	-0.11344186
OPS	0.4081402	0.6344699	0.77465868	0.949926654	1.00000000	0.086983986	-0.04894212
SB	0.3811256	0.1690718	0.23936142	-0.002978809	0.08698399	1.000000000	0.66559295
CS	0.3249182	0.0168328	0.09642614	-0.113441857	-0.04894212	0.665592954	1.00000000

Above, some interesting subsections of the correlation coefficients matrix are presented. The underlined coefficients exceed the threshold that has been set for determining a correlation sufficient enough to warrant removal of a variable. As would be suspected, many of the variables have some level of correlation with some variables having a very high level of correlation.

The HR and RBI correlation is quite high, for obvious reasons. If a player is hitting a lot of homeruns, if players are on base, he is driving in a lot of those runs. And considering when a homerun is hit, it automatically counts as an RBI, there is a built-in correlation between the two. Additionally, there is a high correlation between SLG and OPS, which is also a built-in correlation since the OPS quantity is just a sum of OBP and SLG. The reason the correlation between OBP and OPS is less is because OBP is usually 2-3x smaller than SLG and varies much less.



Finally, AB, R, and H.1B all have very high correlation coefficients and this is not at all difficult to believe. A player with many at bats will also have many hits since they have had many more opportunities to have hits than other players. By the same token, a player with many at bats will also score many runs. Additionally, because most leadoff hitters have the most at bat, they also score the most runs because that is their specialty. As shown, AB has a high level of correlation with all the variables, which makes sense since more attempts would yield more results, for the most part.



B. Resolving Multicollinearity Issues (Somewhat)

For the first set of 3 highly correlated variables, R, AB, and H.1B a regression on each individual variable and each combination of the variables was run.

Var Used	R^2 Value	Reg P-val
R	0.4201	6.98E-16
AB	0.3136	2.00E-11
H.1B	0.3816	3.49E-14
R+AB	0.4026	7.91E-15
R+H.1B	0.4271	4.05E-15
AB+H.1B	0.4014	4.18E-14
R+AB+H.1B	0.4575	1.27E-15

As shown in the table above, as a single variable, R is the most useful as a predictor for the independent variable. Adding H.1B to the model, in addition to R only adds slightly to the R^2 value of the regression and isn't really worth the trade-off in model complexity. Additionally, there is a debatable difference in the effectiveness of the regression relation if R+AB+H.1B is used as opposed to just R. However, to keep model complexity manageable *and* because the P-value for R itself is smaller, AB and H.1B will be dropped from the model.

Next, for HR and RBI the decision is much clearer. RBI is a better predictor variable than HR alone and using both HR and RBI does not in any significant way, increase the predictive power of the model. As such, HR will be dropped from the model.

Var Used	R^2 Value	Reg P-val
HR	0.3309	4.22E-12
RBI	0.3941	1.02E-14
HR+RBI	0.3978	7.81E-14

Finally, for OPS and SLG, it is shown by the table below that both variables have very similar effectiveness in predicting Y. However, since the combination of both is not significantly better than either alone, and since SLG is better alone, OPS will be dropped from the model.

Var Used	R^2 Value	Reg P-val
OPS	0.3094	2.89E-11
SLG	0.3107	2.59E-11
OPS+SLG	0.3179	1.29E-10

Of the original 15 independent variables, AB, H.1B, HR (surprisingly), and OPS have been eliminated, still leaving 11. At first glance, the fact that HR was eliminated may seem strange, since everyone knows that big home run hitters are paid very well in major league baseball. But after some thinking, it is easy to convince one's self that RBI's encompass most of the predictive power of HR's, and then some. The elimination of the

other variables is also intuitive, since much of the discussion of the reason for their correlation is based on common sense of the game of baseball.

Unfortunately, beyond the eliminated variables, many other variables in this data set have high levels of correlation with each other. This is very difficult to deal with, as many do not meet the threshold for elimination. This characteristic of the data will make conclusions difficult to make once the regression is completed.

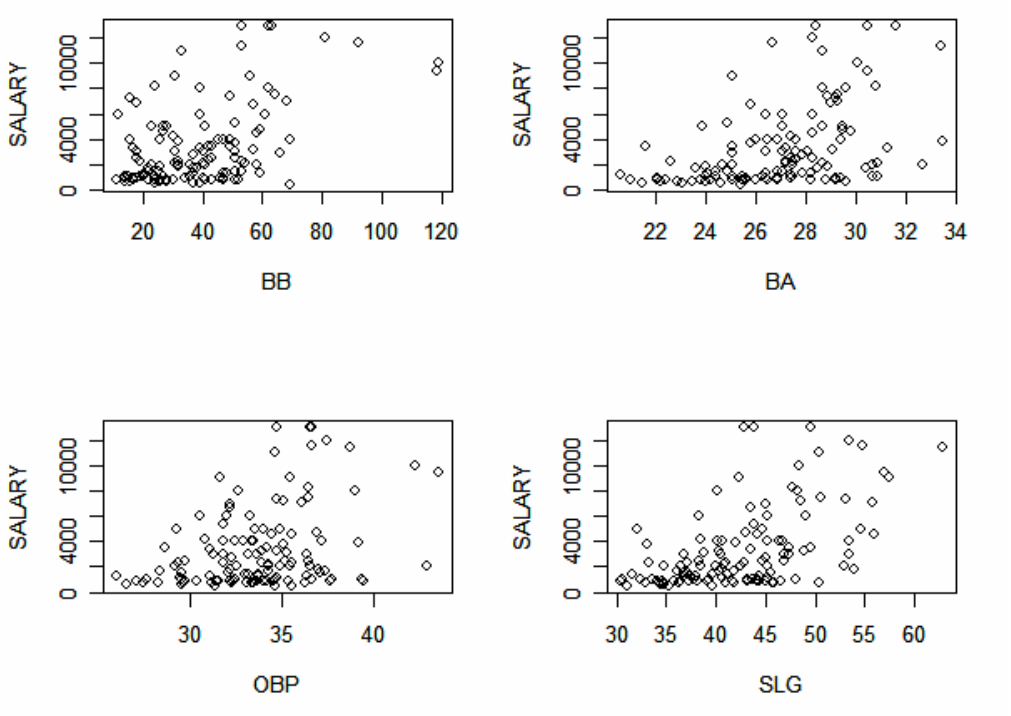
III. Model Creation

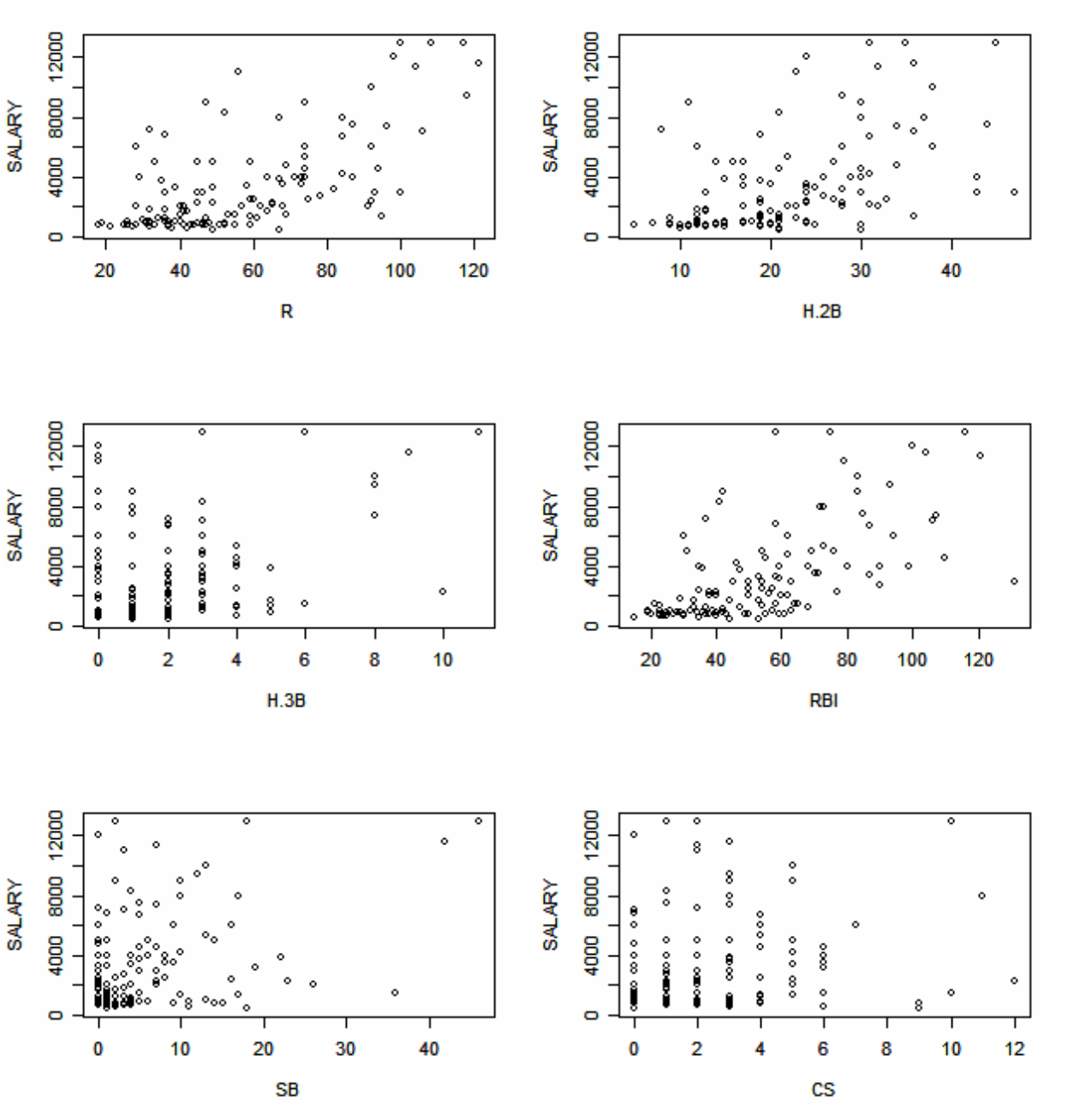
A. Order of the Terms

The next step is to examine the plots of each independent variable versus the dependent variable to see if the correct polynomial order of the term (for use in the model) can be determined.

For R, RBI, and BA, the relationship seems to be relatively parabolic so quadratic terms for these variables will be used in the model. For H.2B, SB, and SLG, there seem to be hints of a quadratic relationship, so to be safe, quadratic terms for these variables will be used as well.

For the other variables, it is difficult to see any relation beyond a linear one (or to see a relation at all), so there will be no quadratic terms for these variables. This will add 6 new terms to the full model, R^2, RBI^2, BA^2, $H.2B^2$, SB^2, and SLG^2.

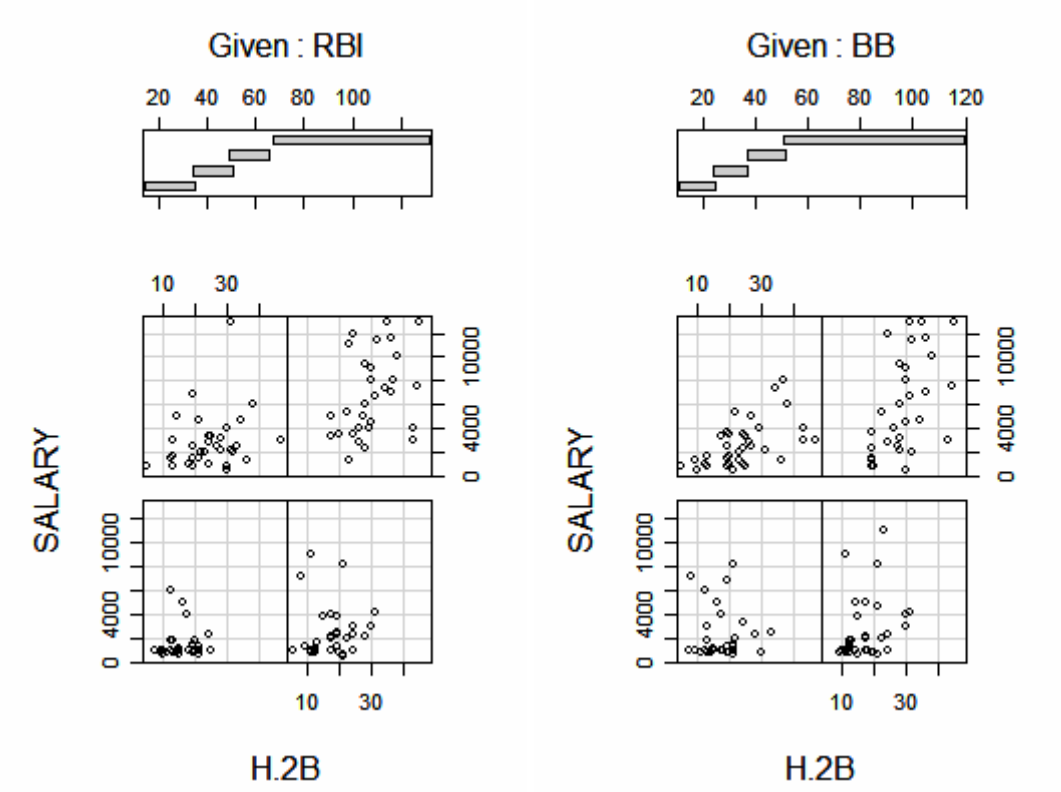


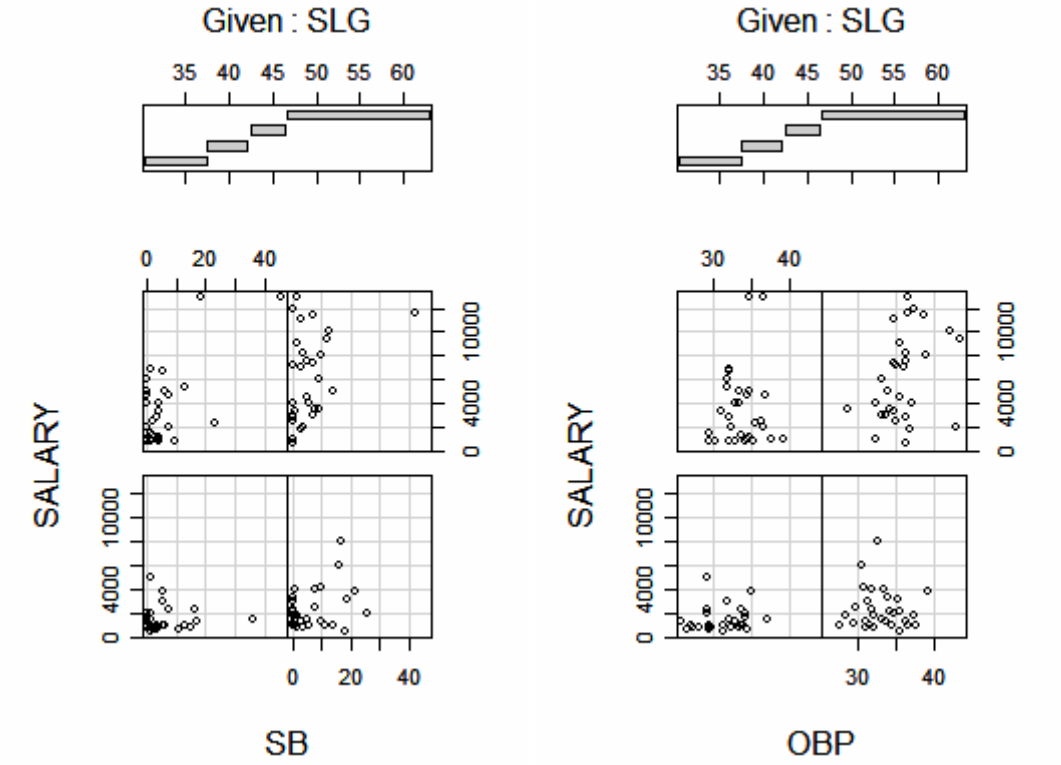


B. Interaction Terms

The next step is to find interactions between the independent variables to see what interaction terms should be used in the full model. Of the 10 choose 2 (50) possible interactions here, only 4 of them are significant and many of them are difficult to interpret and/or very minor.

The 4 relatively significant interactions will be used in this model, H.2B*RBI, H.2B*BB, SB*SLG, and OBP*SLG. All other very minor interactions are left out to keep the model from becoming unwieldy.


Given : RBI
20 40 60 80 100
10 30
0 4000 10000
SALARY
10 30
H.2B
Given : BB
20 40 60 80 100 120
10 30
0 4000 10000
SALARY
10 30
H.2B


Given : SLG
35 40 45 50 55 60
0 20 40
0 4000 10000
SALARY
0 20 40
SB
Given : SLG
35 40 45 50 55 60
30 40
0 4000 10000
SALARY
30 40
OBP

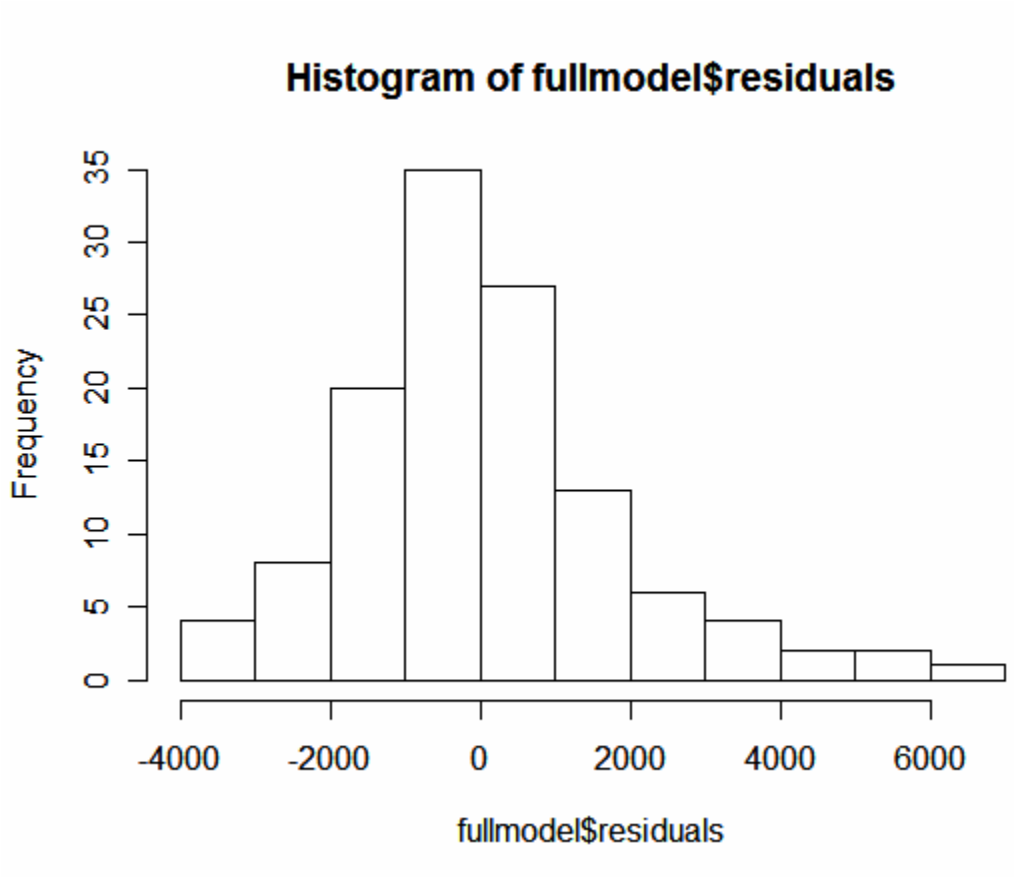
C. Full Model

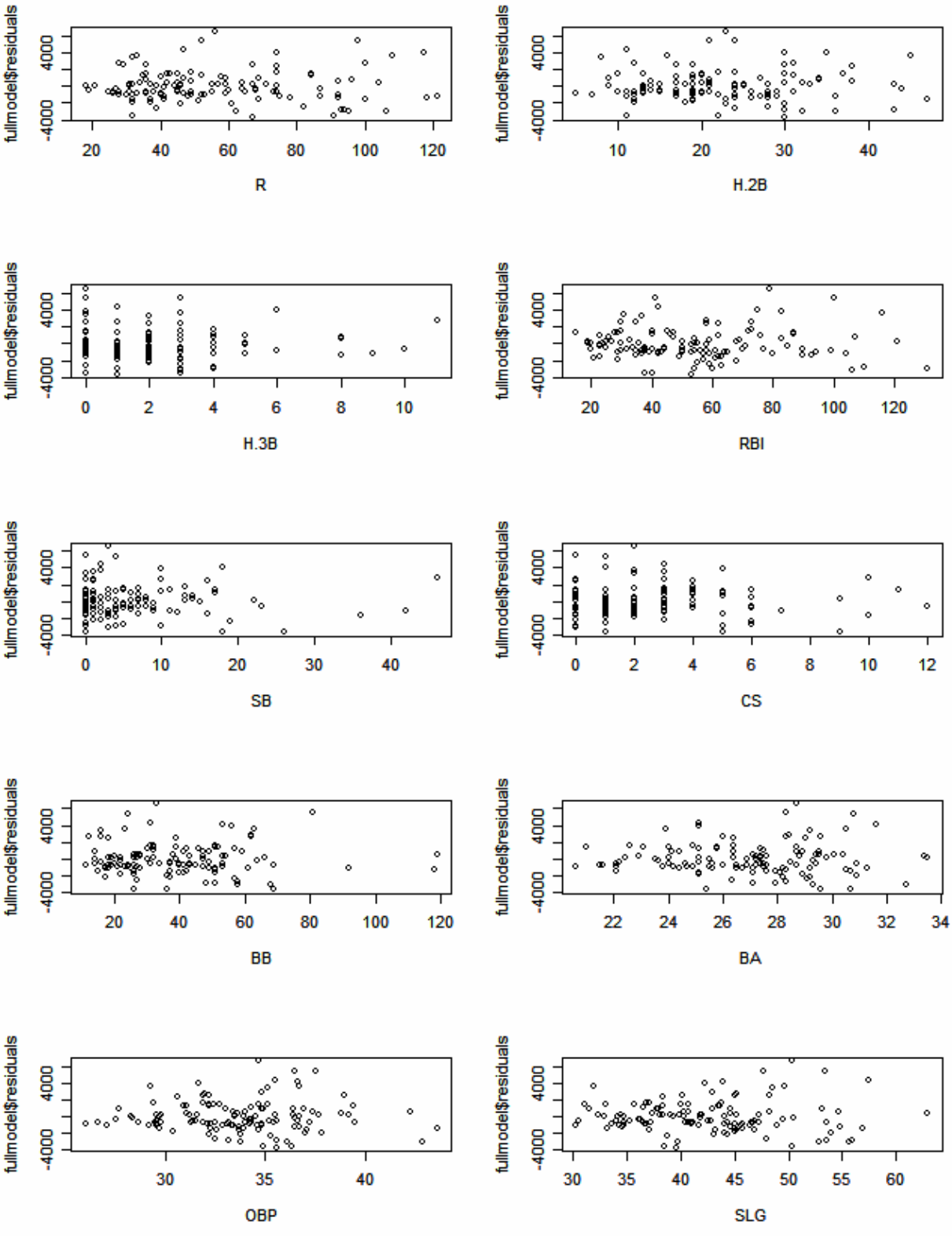
After determining dependent variables to use, the polynomial levels of the terms, and the interaction terms, the full model is show below.

$$Y = B_0 + B_1(R) + B_2(H.2B) + B_3(H.3B) + B_4(RBI) + B_5(SB) + B_6(CS) + B_7(BB) + B_8(BA) + B_9(OBP) + B_{10}(SLG) + B_{11}(P.1B) + B_{12}(P.2B) + B_{13}(P.3B) + B_{14}(P.SS) + B_{15}(P.C) + B_{16}(R^2) + B_{17}(H.2B^2) + B_{18}(RBI^2) + B_{19}(SB^2) + B_{20}(BA^2) + B_{21}(SLG^2) + B_{22}(H.2B*RBI) + B_{23}(H.2B*BB) + B_{24}(SB*SLG) + B_{25}(OBP*SLG) + E$$

IV. Assumption Analysis

All the independent variables are plotted against the residuals to confirm that all the assumptions that are necessary to run a regression, hold. All the variables look to have a random scatter of points and this shows that the variance of the points is relatively constant. As shown below, the residuals look to have a standard distribution.




fullmodel$residuals
-4000
4000
20
40
60
80
100
120
R
10
20
30
40
H.2B
0
2
4
6
8
10
H.3B
20
40
60
80
100
120
RBI
0
10
20
30
40
SB
0
2
4
6
8
10
12
CS
20
40
60
80
100
120
BB
22
24
26
28
30
32
34
BA
30
35
40
OBP
30
35
40
45
50
55
60
SLG

V. Backward Elimination

A. The Elimination Process

Backward elimination will be used to find a much simpler model that seeks to maintain as much predictive power as possible. The alpha value used will be a = .10.

Step #	R^2 Value	Adj R^2	Reg. p-val	Var Removed	p-val Removed
0	0.6503	0.5592	6.40E-13	None	None
1	0.6503	0.5638	2.29E-13	P.2B	0.9807
2	0.6503	0.5682	7.96E-14	P.SS	0.9686
3	0.6502	0.5725	2.71E-14	P.3B	0.9051
4	0.6501	0.5766	9.07E-15	OBP*SLG	0.8489
5	0.6491	0.5796	3.30E-15	P.1B	0.5941
6	0.6480	0.5825	1.19E-15	BA^2	0.5763
7	0.6454	0.5834	5.08E-16	P.C	0.3844
8	0.6426	0.5841	2.20E-16	SB*SLG	0.3663
9	0.6384	0.5832	2.20E-16	H.2B*BB	0.2715
10	0.6330	0.5811	2.20E-16	SB^2	0.2167
11	0.6295	0.5810	2.20E-16	CS	0.3113
12	0.6256	0.5805	2.20E-16	H.3B	0.2944
13	0.6214	0.5797	2.20E-16	$H.2B^2$	0.2719

After the elimination process concludes, this leaves ReducedModel1:

Y = $B_0 + B_1(R) + B_1(H.2B) + B_2(RBI) + B_3(SB) + B_4(BB) + B_5(BA) + B_6(OBP) + B_7(SLG) + B_8(R^2) + B_9(RBI^2) + B_{10}(SLG^2) + B_{11}(H.2B*RBI) + E$

```
Call:
lm(formula = SALARY ~ R + H.2B + RBI + SB + BB + BA + OBP + SLG +
I(R^2) + I(RBI^2) + I(SLG^2) + H.2B * RBI)

Residuals:
    Min      1Q  Median      3Q     Max
-3456.1 -1105.2  -400.7   747.9  6630.6

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 10555.8460  7589.4240   1.391  0.16710
R            -103.8595    48.5169  -2.141  0.03453 *
H.2B         -152.3628    98.2844  -1.550  0.12399
RBI            96.5492    45.1173   2.140  0.03459 *
SB             77.2310    32.1523   2.402  0.01799 *
BB             58.3822    30.8024   1.895  0.06069 .
BA            511.6991   188.8497   2.710  0.00783 **
OBP          -437.5405   202.6297  -2.159  0.03302 *
SLG          -502.3347   367.1665  -1.368  0.17408
I(R^2)          0.8389     0.3610   2.324  0.02199 *
I(RBI^2)       -1.1045     0.4756  -2.322  0.02207 *
I(SLG^2)        7.3621     4.1820   1.760  0.08114 .
```

```
H.2B:RBI        2.6001     1.5339   1.695  0.09292 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2056 on 109 degrees of freedom
Multiple R-Squared: 0.6214,     Adjusted R-squared: 0.5797
F-statistic: 14.91 on 12 and 109 DF,  p-value: < 2.2e-16
```

All of the model terms have p-values < .1, except SLG and H.2B which are used in interaction and polynomial terms. The R-squared value of .6214 is still quite close to the original value of .6503 and more than half of the terms have been removed to make the model much more compact, and manageable.

B. A Simpler Model (A Further Reduction)

Here, a continuation of the backward elimination process is done, with a = .08. This allows a further reduced model with a slightly smaller R-squared value, but allows interpretation to be much easier and allows stronger conclusions to be drawn.

Step #	R^2 Value	Adj R^2	Reg. p-val	Var Removed	p-val Removed
14	0.6114	0.5725	2.2E-16	H.2B*RBI	0.0929
15	0.6114	0.5764	2.2E-16	H.2B	0.9288
16	0.6035	0.5716	2.2E-16	SLG^2	0.1362
17	0.6000	0.5717	2.2E-16	RBI^2	0.3230
18	0.5947	0.5698	2.2E-16	RBI	0.2222

After the elimination process concludes, this leaves ReducedModel2:

$$Y = B_0 + B_1(R) + B_2(SB) + B_3(BB) + B_4(BA) + B_5(OBP) + B_6(SLG) + B_7(R^2) + E$$

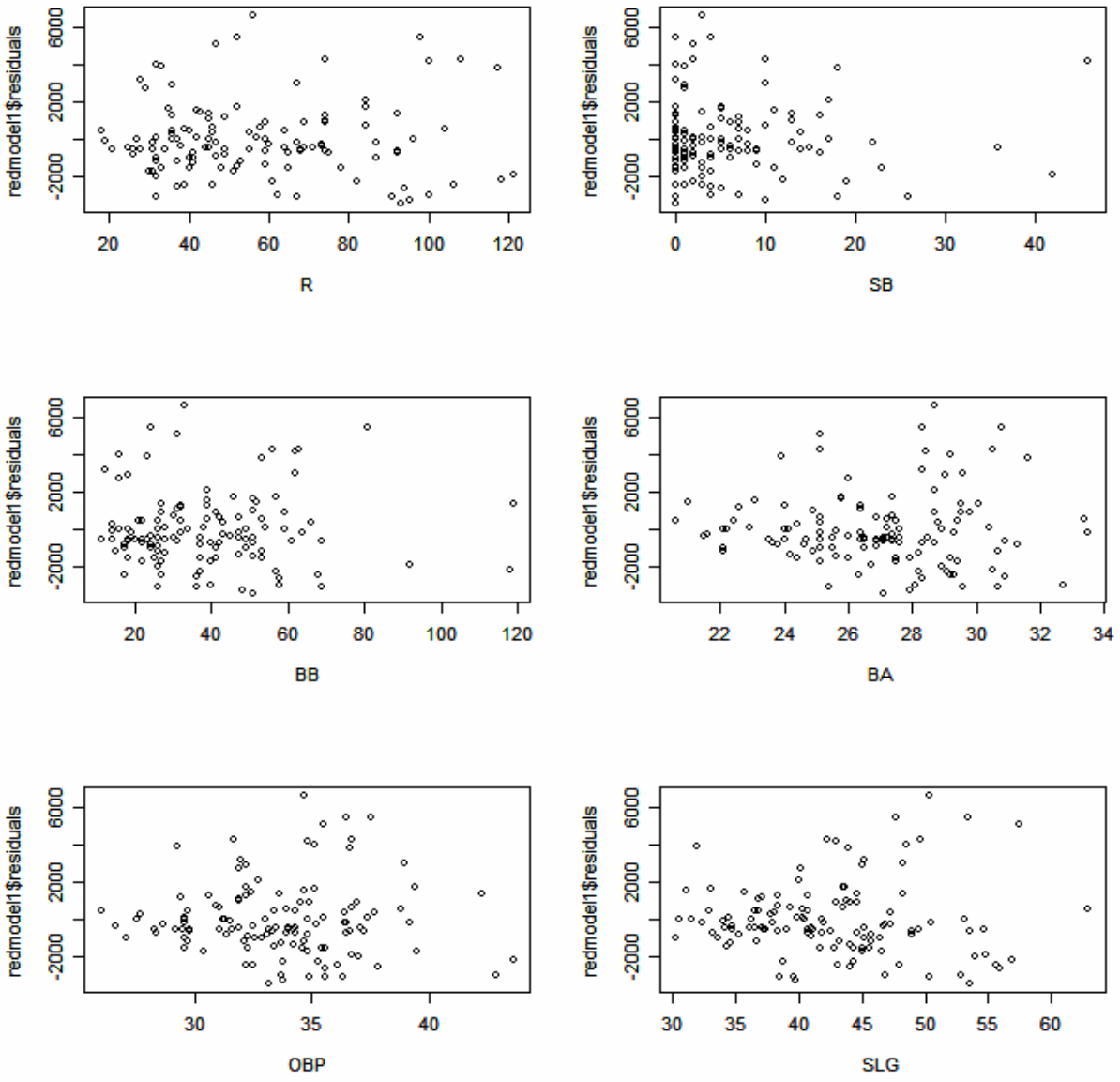
```
Call:
lm(formula = SALARY ~ R + SB + BB + BA + OBP + SLG + I(R^2))

Residuals:
    Min      1Q  Median      3Q     Max
-3751.5 -1160.7  -412.0   799.5  6920.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2504.7325  2470.0363  -1.014 0.312708
R            -103.9077    41.0100  -2.534 0.012644 *
SB             60.3053    29.6807   2.032 0.044499 *
BB             78.6860    26.2170   3.001 0.003303 **
BA            589.4273   168.1366   3.506 0.000652 ***
OBP          -550.7921   166.7498  -3.303 0.001278 **
SLG           179.9167    39.6015   4.543 1.39e-05 ***
I(R^2)          0.8780     0.3052   2.876 0.004801 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2080 on 114 degrees of freedom
```

```
Multiple R-Squared: 0.5947,     Adjusted R-squared: 0.5698
F-statistic: 23.89 on 7 and 114 DF,  p-value: < 2.2e-16
```

All of the model terms have p-values < .045. The R-squared value .5947 is only 9% less than the original value of .6503 and beyond ReducedModel1, this model is much easier to interpret due to the removal of polynomial and interaction terms.

C. Final Assumptions Analysis

For completeness, there must be a check that the residuals of the reduced regression models are still under the assumptions that are required. As shown, the plots below still have a relatively random scatter, showing that the assumptions still hold.



VI. Conclusions – Based on Subset Analyses

A. Power Hitting

From ReducedModel1 to ReducedModel2, all the eliminated variables (H.2B*RBI, H.2B, SLG^2, RBI^2, and RBI) represent statistics that are highly correlated with a player's power hitting. As such, the first order of business is checking to see whether ReducedModel2 appropriately compensates power hitters. The only term left in ReducedModel2 that has a clear relation to power hitters is the SLG term. This must have a heavy weight in the model to appropriately compensate power hitters. Indeed, the large SLG coefficient paired with the smallest p-value on the chart (1.39E095) shows that SLG is indeed very important to the regression relation.

Running the regression on SLG alone produces some interesting results. The R-squared value in this regression (shown below) is .3107. This shows that this regression's effectiveness is approximately 52% of the effectiveness of the *entire* ReducedModel2. Since it was determined earlier that SLG must be used heavily in this model, this finding is expected.

```
Call:
lm(formula = SALARY ~ SLG)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -8020.37    1564.04  -5.128 1.14e-06 ***
SLG           267.06      36.31   7.354 2.59e-11 ***
---

Residual standard error: 2644 on 120 degrees of freedom
Multiple R-Squared: 0.3107,     Adjusted R-squared: 0.3049
F-statistic: 54.08 on 1 and 120 DF,  p-value: 2.59e-11
```

If the correlation matrix for the model variables is examined, it is seen that SLG has a correlation coefficient < .53 with all of the other model variables. This shows that a major part of the model's power comes from this single variable. It can be concluded, that indeed a player's slugging percentage is a good predictor for a player's salary.

```
             R            SB        BB        BA       OBP
SLG  0.4799744 -0.002978809 0.4093818 0.4796262 0.5383092
```

Finally, if the correlation matrix between SLG and HR, and RBI is viewed, one can easily see that much of the variance captured by HR and RBI is also captured by SLG which makes it a little more legitimate that HR and RBI are not included in ReducedModel2.

```
          SLG       HR       RBI
SLG 1.0000000 0.8098600 0.6886996
HR  0.8098600 1.0000000 0.8672076
RBI 0.6886996 0.8672076 1.0000000
```

B. Hitting for Average

In addition to clearing the bases, driving in runs, and hitting a lot of homeruns, players that get on base, have plate discipline, and can put the ball into play are big assets to any baseball team. The main statistics that capture these qualities are BB, BA, and OBP.

```
Call:
lm(formula = SALARY ~ BB + OBP + BA)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3719.12    2592.13  -1.435    0.154
BB            121.72      16.42   7.413 2.05e-11 ***
OBP          -612.25     140.75  -4.350 2.91e-05 ***
BA            852.38     136.77   6.232 7.36e-09 ***
---

Residual standard error: 2394 on 118 degrees of freedom
Multiple R-Squared: 0.4444,     Adjusted R-squared: 0.4303
F-statistic: 31.46 on 3 and 118 DF,  p-value: 5.127e-15
```

Combined, these 3 statistics are less than half the terms in the model, yet they account for almost 75% of ReducedModel2's R^2 value. This may seem like a contradiction since SLG accounts for 52% of the model's R^2 value, but the high levels of multicollinearity ensure that there is much overlap between what these statistics measure.

Seperately BB, OBP, and BA yield R-squared values of .2572, .1385, and .1778 which is not entirely unreasonable. OBP and BA are both percentage measures and do not really take into account how many games a player has played. This may mean that a player who has only played a few dozen games will have a good batting average, but that does not mean they will earn a significant amount of money. However, if a player has a large number of walks it means that not only do they have plate discipline, but pitchers also respect their hitting, *and* they have had enough at bats to tally up the walks. This explains why walks are a very important factor when determining a player's plate efficiency.

C. Speed

Finally, the last bit of player performance that was hypothesized to affect player salary is their speed. Being much less of a factor than power and batting average (stealing is less pervasive and not used as often), it is expected that steals will have a marginal affect on salary, but not to a great extent.

```
Call:
lm(formula = SALARY ~ SB)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  2585.77     336.33   7.688 4.56e-12 ***
SB            129.39      33.65   3.845 0.000194 ***
---

Residual standard error: 3005 on 120 degrees of freedom
Multiple R-Squared: 0.1097,     Adjusted R-squared: 0.1023
F-statistic: 14.78 on 1 and 120 DF,  p-value: 0.0001943
```

Here, the result is not exactly what one might want to find (especially a fast baseball player.) Stolen bases are (apparently) a relatively poor indicator for a player's salary. This might lead one to believe that steals are not at all important in determining a player's salary. This would be an incorrect assessment. The majority of major league baseball players are paid based on the other two categories (Power and Average) as opposed to speed. As such, this model does not have many players who are speed players and paid for their speed. This leads to the conclusion that it may be better to create two separate equations, for speed players and for the rest of MLB players.

To attempt to get a better feel for how fast players are compensated (or undercompensated, as it has been shown so far), a regression with SB, R, and R^2 will be done. This is because fast players, on average will score more runs than slower players. For example, if a slow player is on 2nd base, a base hit will most likely not score him, but a fast player will score, in all likelihood. This will be done with the caveat that not *all* of the effect of R and R^2 is based on the player's speed. Much of it has to do with the team the player plays for (do they score a lot of runs?), the number of games the player has played, and other factors (homeruns and big hits score runs too).

```
Call:
lm(formula = SALARY ~ SB + R + I(R^2))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2789.5034  1298.6514   2.148 0.033758 *
SB            -4.3147    30.1504  -0.143 0.886449
R            -69.3384    44.0524  -1.574 0.118166
I(R^2)         1.1883     0.3326   3.572 0.000513 ***
---

Residual standard error: 2323 on 118 degrees of freedom
Multiple R-Squared: 0.4768,     Adjusted R-squared: 0.4635
F-statistic: 35.84 on 3 and 118 DF,  p-value: < 2.2e-16
```

The huge jump in the R^2 value for this regression shows how important R and R^2 are as predictor variables in this model. In fact, if a regression is run with R and R^2 alone, the R^2 value turns out to be .4767, a .0001 difference from the SB, R, R^2 regression. This shows that almost all of the variance explained by SB is also explained by R. So why is SB still in the model? As shown earlier, the p-value of SB was .044499, not a relatively low value, but definitely under the determined threshold of elimination.

VII. Overall Conclusions

In this study, it is clear that a player's statistics do provide at least some sort of predictive power when it comes to predicting salary for an upcoming year. This is evidenced by the decent R-squared values in ReducedModel1 (0.6214) and ReducedModel2 (.5947). Additionally, both models have minuscule p-values, approaching 0. The majority of the mathematical analysis will be done on ReducedModel2, since the interpretation difficulties due to interaction terms and multicollinearity are beyond the scope of this project.

A. ReducedModel1

With this model, it is very difficult to interpret many of the terms in the model since they are affected by interactions and polynomial terms as well. The linear terms in this model that do not have interaction terms or higher order terms are SB, BB, BA, and OBP. Unfortunately, due to the high multicollinearity between the variables at the onset, even the conclusions to be made about these variables will be imprecise.

Roughly speaking, this model predicts that for every base that a baseball player steals, if he is receiving a new contract the following year, he will receive $77,230 on his salary. For every walk that he is issued, he will receive $58,382 and every batting average point earns him $51,196. The OBP variable has a negative coefficient *not* because a lower OBP is better, but instead because OBP has a high correlation with other variables and as a whole, the its coefficient along with the others give a better estimate of how a player's salary will be affected.

The efficiency and correctness of this model can be viewed through a practical lens as well. Baseball experts know that the major parts of a player's game include their defense, ability to hit for average, ability to hit for power, and their speed. The model below includes aspects of all of these factors (except defense, which is difficult to capture statistically). In terms of ability to hit for average, BB, BA, OBP, and SLG, and SLG^2 all take into account a hitter's plate discipline and propensity to swing at bad pitches, and their ability to put the ball into play. With respect to power, H.2B, SLG, RBI, and RBI^2 all represent the player's ability to put the ball into the outfield, the ability to drive in runs, and the ability to help clear the bases. Finally, with respect to speed, SB, R, and R^2 do a good job of showing how fast a player is, both with and without respect to their ability to get on base. Simply speaking, the appearance of each of these terms is not surprising and easily explainable. The exact interpretation of each of the coefficients requires much more analysis.

Taking an example from this past year, Reggie Sanders:

R = 49, H.2B = 14, RBI = 54, SB = 14, BB = 28, BA = 27.1, OBP = 34.0, SLG = 54.6
R^2 = 2401, RBI^2 = 2916, SLG^2 = 2981.16, H.2B*RBI = 756

Plugging this into the regression model,

Y = 10555.8460 + -103.8595(49) + -152.3628(14) + 96.5492(54) + 77.231(14) + 58.3822(28) + 511.6991(27.1) + -437.5405(34.0) + -502.3347(54.6) + 0.8389(2401) + -1.1045(2916) + 7.3621(2981.16) + 2.6001 (756)

Y = 5533.188 => $5,533,188 is Reggie Sanders predicted salary. His actual salary is $5,000,000.

B. ReducedModel2

The following table shows what percentage of ReducedModel2's R^2 value is explained by the individual regression terms.

Variable	R^2 Value	%
R	0.4201	70.64
SB	0.1097	18.45
BB	0.2572	43.25
BA	0.1778	29.90
OBP	0.1385	23.29
SLG	0.3107	52.24
R+R^2	0.4767	80.16

Due to the high levels of multicollinearity, these percentages do not add up to 100%. This shows that in general R and SLG offer very good predictive power in determining a player's salary for an upcoming year. As explained above, SLG is an excellent indicator of power and an above average indicator of average. R is an average indicator of speed and average indicator of average. Together, these two statistics do indeed cover a lot of the bases, so to speak.

```
Call:
lm(formula = SALARY ~ R + I(R^2) + SLG)

Residuals:
    Min      1Q  Median      3Q     Max
-4426.4 -1363.3  -395.1   817.6  7193.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2654.8679  1788.6284  -1.484 0.140395
R             -71.0657    41.1153  -1.728 0.086524 .
I(R^2)          1.0539     0.3117   3.381 0.000979 ***
SLG           141.5917    34.1570   4.145 6.42e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2170 on 118 degrees of freedom
Multiple R-Squared: 0.5432,     Adjusted R-squared: 0.5316
F-statistic: 46.77 on 3 and 118 DF,  p-value: < 2.2e-16
```

As expected, this combined regression shows a fantastic model. This is a model with *only* R, R^2, and SLG explains 91.3% of the variance explained by the entirety of ReducedModel2 and has less than half as many terms.

C. Final Thoughts

This study has uncovered a multitude of information. The study has shown (as expected) that there is a definite relation between a player's statistical performance and the salary that they receive. This is shown by multiple regressions with relatively good R-squared values. Additionally, specific statistics were found to be of particular importance in terms of predicting salary, namely Slugging Percentage (SLG), Walks (BB), and Runs Scored (R).

Another important "conclusion" that can be drawn is that from the beginning, it was apparent that many baseball statistics had a high level of correlation. All the statistics that were expected to be correlated (RBI and HR, etc) were and statistics that one might not expect to be correlated, were correlated as well. A major reason for this is the fact that as a player has more at bats, other statistics are bound to increase. Even if a player is terrible at driving in runs, given enough at bats, he will have a high RBI count. This accounts for a large portion of the correlation between seemingly uncorrelated statistics. This high level of multicollinearity, as explained in section VII.A, makes it very difficult to draw strong conclusions based on the coefficients of the final regression model.

Another piece of information that may be quite relevant is a surprising conclusion. At the onset of the study, dummy variables representing a player's position were added to the data because it was a possibility that a player's position would affect his salary. Surprisingly, early in the elimination phase, all the player dummy variables were eliminated. This indicates that a player's position is not very important in determining his salary. The only possibly significant exception is for catchers. The P.C variable was not eliminated until the 7^{th} step of the elimination process. This indicates that there may be some modifier on a catcher's salary with respect to player's of other positions.

Finally, it should be noted that through this study, it was found that it may be prudent to have 2 separate models for players that can be classified as "speed" players and others that can be classified as "power" players. This is because some players are paid for their speed and others are paid for their power. When a player has average speed and average power, the effect on their salary is not necessarily additive. This is a pronounced difference since in ReducedModel2 and ReducedModel1, there is relatively little weight given to Stolen Bases. The reason this does *not* indicate that speed is *not* important is because in some of the observations, players with very little power and a lot of speed had their salary predicted quite incorrectly based on the model. The fact that the majority of the players were power players may have influenced the model to make it better suited to predict the salaries of power players.