Predicting America's Next MLB All-Star

Matthew Hoth, Philip Middleton, Siddant Mehta

December 9, 2019

1 The Motivation

Our group loves sports, we were able to bond right off the bat because of our love of sports. All of us watch sports in our personal lives and have started to see how important statistics and data analytics are becoming to sports. For each of us this was a part of the reason to do this program is to learn about the analytical decisions made in sports. Baseball was one of the first major sports to highly implement analytics into team decisions. This was documented in a movie called "Moneyball" and peaked our interest in how analytics created success for a team. Statistics have always been a critical aspect in baseball. The history of baseball is all tracked through statistics and has always played a role in the sport but with modern day analytical tools all these statistics are making very important decisions for a team. Decisions like draft picks, who to play for the game, how much a certain player gets paid, etc. Analytics has a position on every team in the MLB.

With all of these statistics what did we want to discover as a group? Asking ourselves this questions we came to the conclusion that we wanted to predict All-Stars in On-base Percentage Slugging (OPS), Home Runs (HR), and Stolen Bases (SB). Finding these numbers will help us predict future players potential success in all of these categories. All teams want to win games and our analytics can find the numbers that could help teams find their next All-Star.

2 The Data

Our dataset consists of over 14,000 observations across 35 variables. The dataset contains baseball metrics from the 1985 Season to 2016 Season. Moreover, the dataset used for analysis only contains hitting metrics, with the major of pitchers removed from the data. This was done to strengthen the authenticity of our predictions.

Qualitative Variables: Among the 35 variables, there are 5 qualitative variables. Out of these 5 variables, our group will be using variable *All-Star* in order to predict if a player would become an *All-Star* or not. This variable is represented as a dummy variable, with 1 denoting that the player was an All-Star and 0 denoting they weren't. The other qualitative variables in the dataset are the following;*bats, throws, League*, and *Power.Index*. The variable *Power.Index* ranks a players batting power on a scale of four levels; Poor, Average, Great, Excellent. This variable is derived from from the *ISO* variable.

Quantitative Variables: Among the 35, there are 25 quantitative variables. All of these variables are continuous and were used in predicting the following numerical variables; *Home Runs, OPS*, and *Stolen Bases*. A statistical summary of the data is seen in **Figure 1**.

	Mean	Sum	Min	Max	Median	Length	Standard Dev.
G	94.69	1394032.00	14.00	163.00	97.00	14722	43.78
AB	304.55	4483576.00	50.00	716.00	284.00	14722	183.95
R	41.37	609077.00	0.00	152.00	35.00	14722	30.72
H	80.88	1190784.00	1.00	262.00	72.00	14722	54.89
AVG	0.25	3660.90	0.02	0.43	0.26	14722	0.05
OBP	0.31	4617.13	0.03	0.61	0.32	14722	0.06
SLG	0.46	6791.61	0.02	1.15	0.47	14722	0.13
OPS	0.77	11408.61	0.05	1.62	0.79	14722	0.18
ISO	0.21	3130.71	0.00	0.78	0.21	14722	0.10
X2B	15.63	230094.00	0.00	59.00	14.00	14722	11.62
X3B	1.73	25483.00	0.00	23.00	1.00	14722	2.24
HR	9.12	134295.00	0.00	73.00	6.00	14722	9.78
RBI	39.51	581633.00	0.00	165.00	32.00	14722	31.04
SB	5.93	87328.00	0.00	110.00	2.00	14722	9.55
BB	29.98	441298.00	0.00	232.00	24.00	14722	24.49
SO	56.14	826471.00	2.00	223.00	48.00	14722	36.31
salary	2518407.99	37076002376.00	60000.00	33000000.00	775000.00	14722	3863225.70
IBB	2.54	37345.00	0.00	120.00	1.00	14722	4.02
HBP	2.75	40544.00	0.00	35.00	2.00	14722	3.36
SH	2.44	35925.00	0.00	39.00	1.00	14722	3.28
SF	2.60	38223.00	0.00	17.00	2.00	14722	2.49

(1)

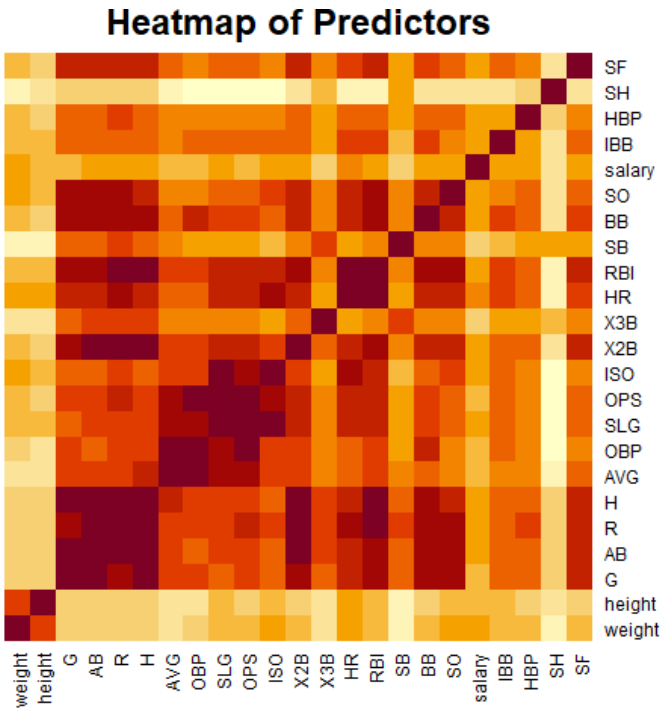
2.1 Data Pre-Processing

Before our group started our analysis on the MLB dataset that we uncovered, we have to remove a few variables from the data. This was done because collinearity existed between some of the variables, and therefore our interpretation of how our predictors impact the outcome wouldn't be precise. The following variables were removed due to collieanerity; *OBP* and *SLG*. Furthermore, since pitchers can be batters, only when they play in the Nation League (NL), they were included in the original dataset. These pitchers would play a max 50 games, depending on whether they were starters or relief pitchers, and statistically batted the worse among all other positions. The data, therefore, contained players who were potential outliers and would skew distribution of varialbes, potentially leading to decreased prediction accuracy. We removed these players, along with the 4 variables mentioned previously, before we started our analysis.

3 Visuals

3.1 Correlations:

Figure 2 shows a heatmap of all of our possible predictors in the dataset. The deeper the red, the more positive the correlation is between the two variables while the light yellow symbolizes a more negative correlation between the two variables. Based on this heatmap, we notice that a lot of our hitting metrics are highly correlated with one another. For instance, G and AB share a very deep red pointing towards them sharing a high positive correlation. In addition to noticing correlations in our data, we used this heatmap to help also find potential collinearity among our variables. As mentioned before, we did end up removing 4 variables due to collinearity, and this heatmap helped our group determine those exact variables. Along with this heatmap, **Figure 3** shows the correlation matrix used to create the heatmap. This matrix displays the same information as the heatmap but it easier to interpret.



heatmap.png

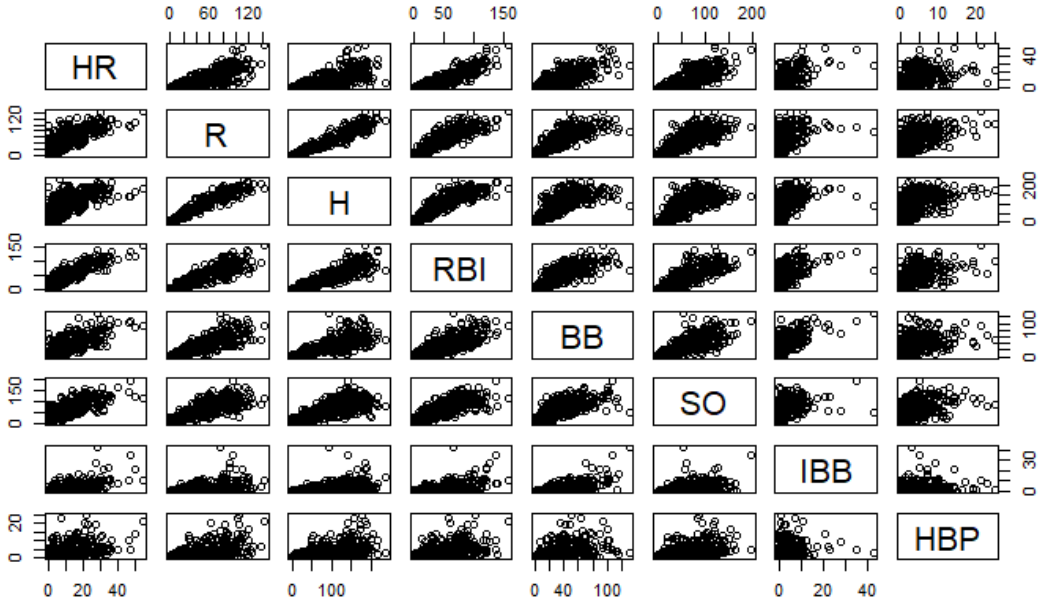
(2)

	weight	height	G	AB	R	H	AVG	OBP	SLG	OPS	ISO	X2B	X3B	HR	RBI	SB	BB	SO	salary	IBB	HBP	SH	SF
weight	1.00	0.53	0.00	0.03	0.03	0.02	-0.09	-0.06	0.14	0.08	0.24	0.09	-0.16	0.26	0.16	-0.20	0.07	0.24	0.25	0.11	0.11	-0.21	0.05
height	0.53	1.00	-0.07	-0.04	-0.03	-0.05	-0.15	-0.12	0.06	0.00	0.16	0.00	-0.16	0.19	0.10	-0.20	0.03	0.14	0.14	0.10	-0.03	-0.13	0.02
G	0.00	-0.07	1.00	0.95	0.88	0.92	0.53	0.53	0.52	0.55	0.43	0.86	0.51	0.67	0.83	0.45	0.77	0.77	0.14	0.48	0.49	-0.01	0.66
AB	0.03	-0.04	0.95	1.00	0.94	0.98	0.55	0.52	0.55	0.57	0.45	0.91	0.54	0.71	0.88	0.49	0.79	0.79	0.22	0.49	0.51	-0.02	0.69
R	0.03	-0.03	0.88	0.94	1.00	0.95	0.60	0.61	0.64	0.66	0.56	0.89	0.55	0.78	0.89	0.55	0.85	0.77	0.24	0.52	0.53	-0.06	0.66
H	0.02	-0.05	0.92	0.98	0.95	1.00	0.65	0.59	0.60	0.63	0.49	0.92	0.55	0.72	0.89	0.50	0.78	0.74	0.22	0.51	0.51	-0.04	0.69
AVG	-0.09	-0.15	0.53	0.55	0.60	0.65	1.00	0.89	0.79	0.86	0.55	0.60	0.35	0.43	0.56	0.30	0.47	0.32	0.09	0.34	0.30	-0.25	0.42
OBP	-0.06	-0.12	0.53	0.52	0.61	0.59	0.89	1.00	0.78	0.90	0.60	0.56	0.29	0.48	0.56	0.28	0.66	0.37	0.11	0.42	0.36	-0.32	0.39
SLG	0.14	0.06	0.52	0.55	0.64	0.60	0.79	0.78	1.00	0.98	0.95	0.65	0.29	0.74	0.71	0.17	0.56	0.51	0.18	0.43	0.36	-0.41	0.45
OPS	0.08	0.00	0.55	0.57	0.66	0.63	0.86	0.90	0.98	1.00	0.88	0.65	0.30	0.69	0.69	0.21	0.62	0.49	0.16	0.45	0.38	-0.40	0.46
ISO	0.24	0.16	0.43	0.45	0.56	0.49	0.55	0.60	0.95	0.88	1.00	0.57	0.21	0.78	0.67	0.07	0.52	0.53	0.19	0.41	0.34	-0.43	0.40
X2B	0.09	0.00	0.86	0.91	0.89	0.92	0.60	0.56	0.65	0.65	0.57	1.00	0.45	0.71	0.87	0.39	0.74	0.72	0.23	0.49	0.50	-0.11	0.67
X3B	-0.16	-0.16	0.51	0.54	0.55	0.55	0.35	0.29	0.29	0.30	0.21	0.45	1.00	0.22	0.37	0.58	0.37	0.38	-0.02	0.17	0.23	0.13	0.30
HR	0.26	0.19	0.67	0.71	0.78	0.72	0.43	0.48	0.74	0.69	0.78	0.71	0.22	1.00	0.90	0.17	0.73	0.76	0.31	0.57	0.46	-0.30	0.57
RBI	0.16	0.10	0.83	0.88	0.89	0.89	0.56	0.56	0.71	0.69	0.67	0.87	0.37	0.90	1.00	0.29	0.79	0.78	0.28	0.59	0.49	-0.21	0.73
SB	-0.20	-0.20	0.45	0.49	0.55	0.50	0.30	0.28	0.17	0.21	0.07	0.39	0.58	0.17	0.29	1.00	0.38	0.33	0.02	0.14	0.22	0.18	0.24
BB	0.07	0.03	0.77	0.79	0.85	0.78	0.47	0.66	0.56	0.62	0.52	0.74	0.37	0.73	0.79	0.38	1.00	0.71	0.26	0.63	0.43	-0.13	0.59
SO	0.24	0.14	0.77	0.79	0.77	0.74	0.32	0.37	0.51	0.49	0.53	0.72	0.38	0.76	0.78	0.33	0.71	1.00	0.22	0.39	0.48	-0.12	0.52
salary	0.25	0.14	0.14	0.22	0.24	0.22	0.09	0.11	0.18	0.16	0.19	0.23	-0.02	0.31	0.28	0.02	0.26	0.22	1.00	0.24	0.17	-0.09	0.18
IBB	0.11	0.10	0.48	0.49	0.52	0.51	0.34	0.42	0.43	0.45	0.41	0.49	0.17	0.57	0.59	0.14	0.63	0.39	0.24	1.00	0.24	-0.19	0.41
HBP	0.11	-0.03	0.49	0.51	0.53	0.51	0.30	0.36	0.36	0.38	0.34	0.50	0.23	0.46	0.49	0.22	0.43	0.48	0.17	0.24	1.00	-0.06	0.35
SH	-0.21	-0.13	-0.01	-0.02	-0.06	-0.04	-0.25	-0.32	-0.41	-0.40	-0.43	-0.11	0.13	-0.30	-0.21	0.18	-0.13	-0.12	-0.09	-0.19	-0.06	1.00	-0.11
SF	0.05	0.02	0.66	0.69	0.66	0.69	0.42	0.39	0.45	0.46	0.40	0.67	0.30	0.57	0.73	0.24	0.59	0.52	0.18	0.41	0.35	-0.11	1.00

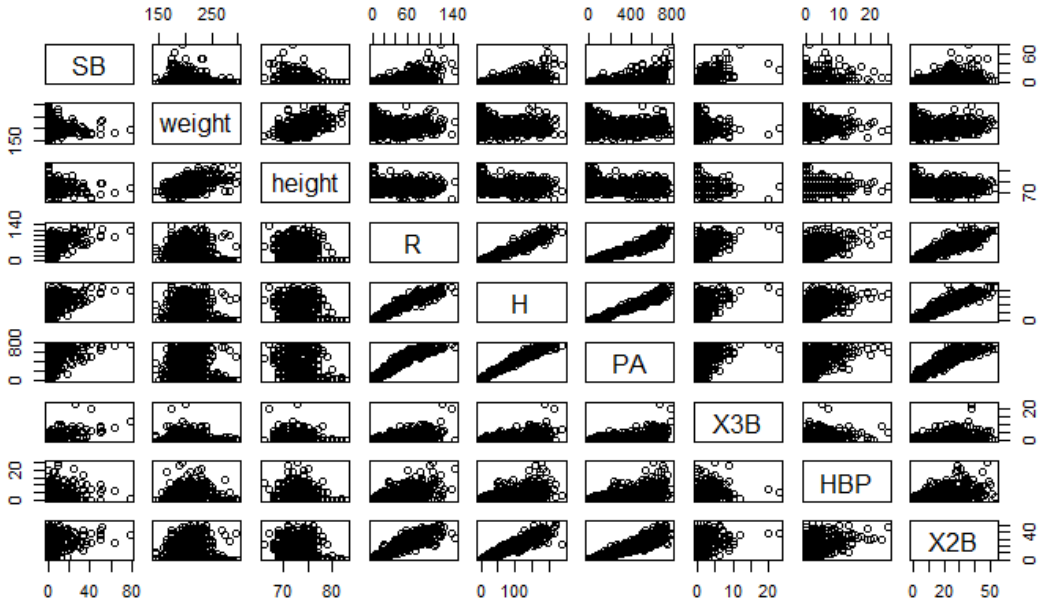
(3)

3.2 Pairwise Plots:

These two pairwise plots, **Figure 4** and **Figure 5**, represents some of the relationship that our predictors have with one another. This plots take our interpretations of correlation a step further by examining if the relationship between variables is linear, non-linear, or collinear. The variables at interest in these two plots are *HR* and *SB* and we used these plots to determine how to adjust the non-linearity in our linear models.



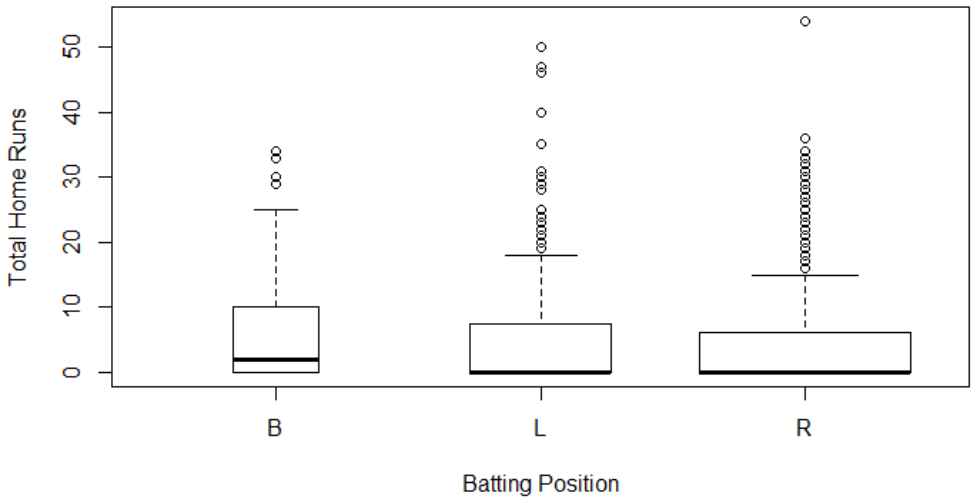
(4)



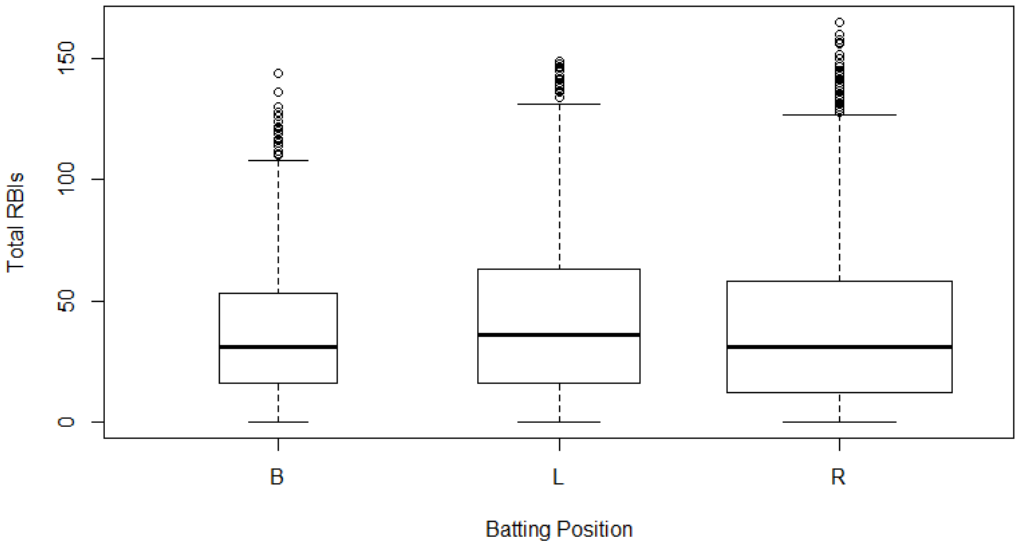
(5)

3.3 Boxplots:

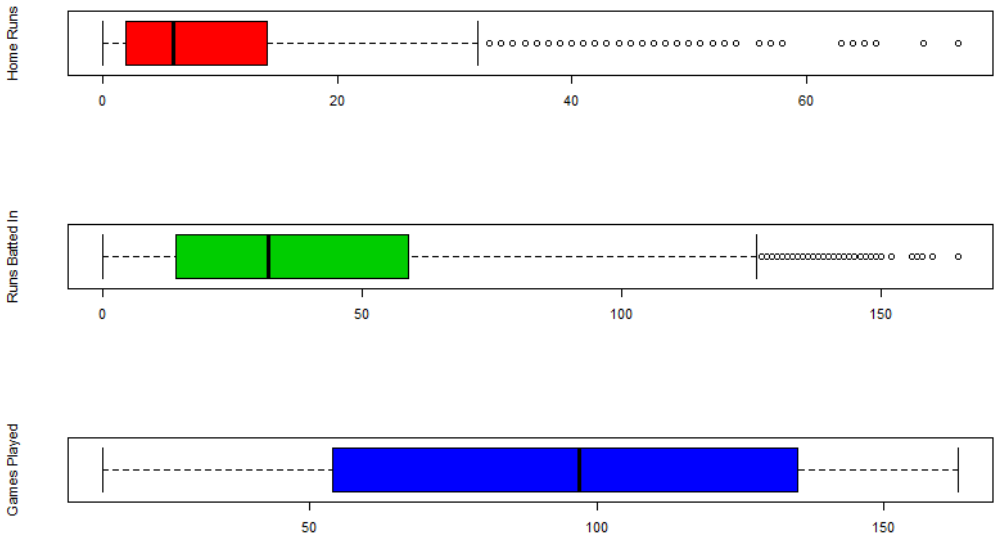
Figures 6 and **Figure 7** display side by side boxplots looking at batting stance on the x-axis, with the y-axis in **Figure 6** displaying Total Home Runs and **Figure 7** displaying Total RBIs. These two side by side boxplots were created to see if there is a relationship between a player's batting stance and player's ability to produce a run. However, our boxplot's indicated that no relationship is visible. **Figure 8** shows three separate boxplots representing variables that are heavily associated with player production, these variables being; HR, RBI, and Games Played (G).



to bat boxplot.png (6)

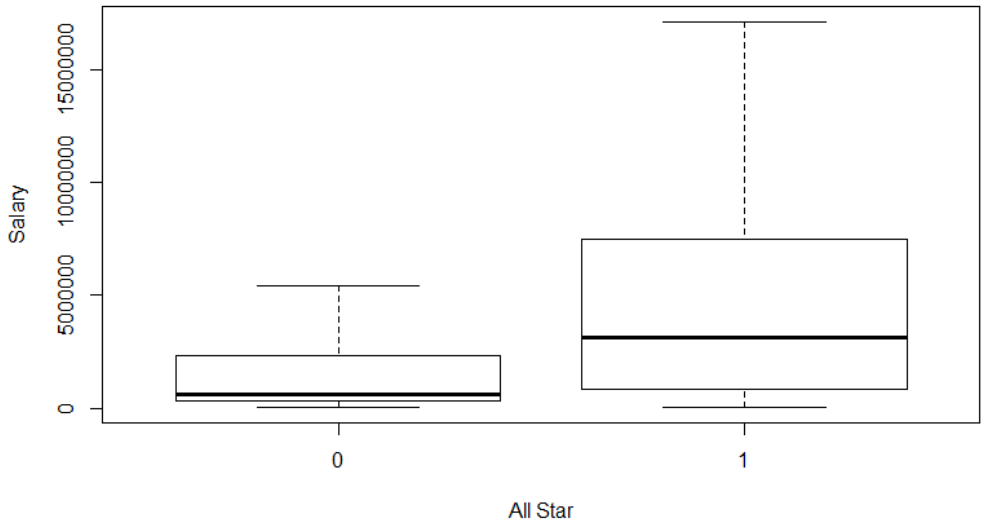


boxplot.png (7)

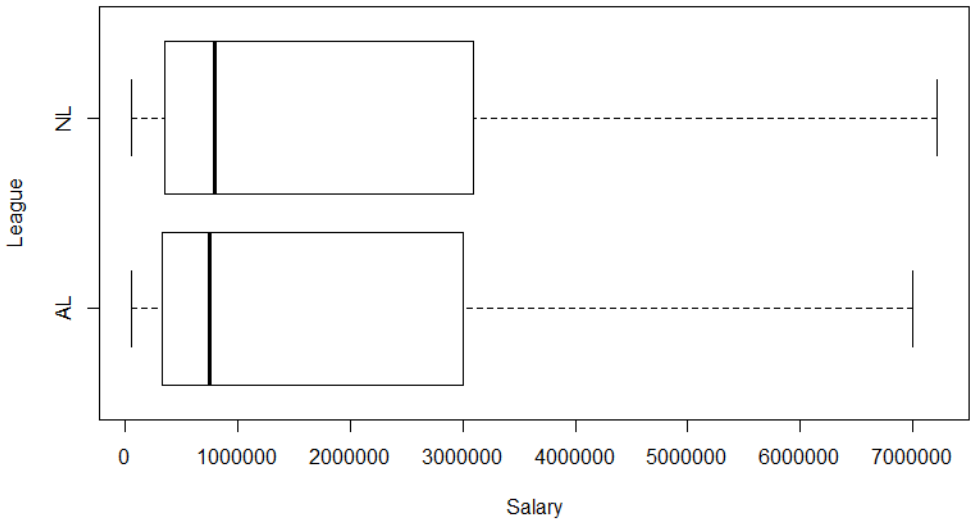


boxplots hr rbi g.png (8)

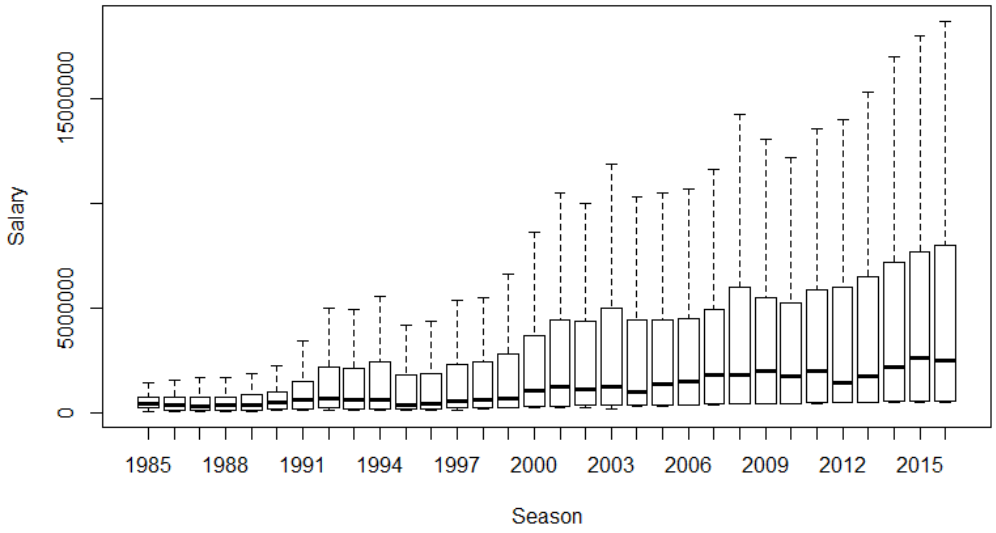
Figure 9 displays a boxplot displaying All-Stars and non All-Stars and their salary. The plot tells us very clearly that All-Star's have a higher median salary than non All-Star's and that the maximum salary offered to an All-Star is significantly larger than a non All-Star. **Figure 10** then shows the salary of players by league. Our group noticed that NL has a slightly higher median salary compared to the AL. Finally, **Figure 11** shows multiple side by side boxplots of *salary* across every season in order. Here we see not only an increase in median salary as *season* increases but, more importantly, we see the maximum value of salary in each boxplot grow exponentially as *season* increases.



star sal boxplot.png (9)



by league boxplot.png (10)



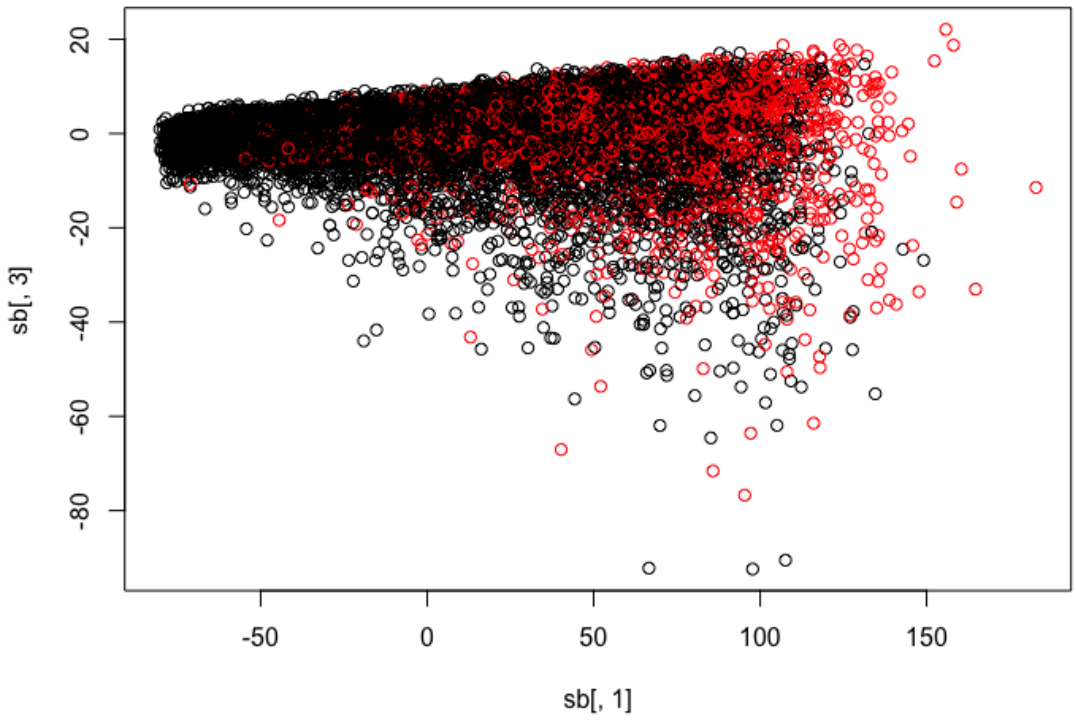
sal by season.png (11)

4 Principal Component Analysis

When we run our PCA we can see that we only need 7 principal components to account for 88% of our data. When looking at the first principal component we can the highest variation is captured most by runs, hits, at bats, and RBIs. In the second principal component we can determine that the highest variation is in batting average, on-base percentage, slugging, and on-base percentage slugging. The third principal component captures most of the height and weight variables. In our fourth principal component we see the highest measurement is in sacrifice hits, height and weight. The fifth principal component accounts for mainly height and also has a higher variation in weight and triples. When looking at the sixth principal

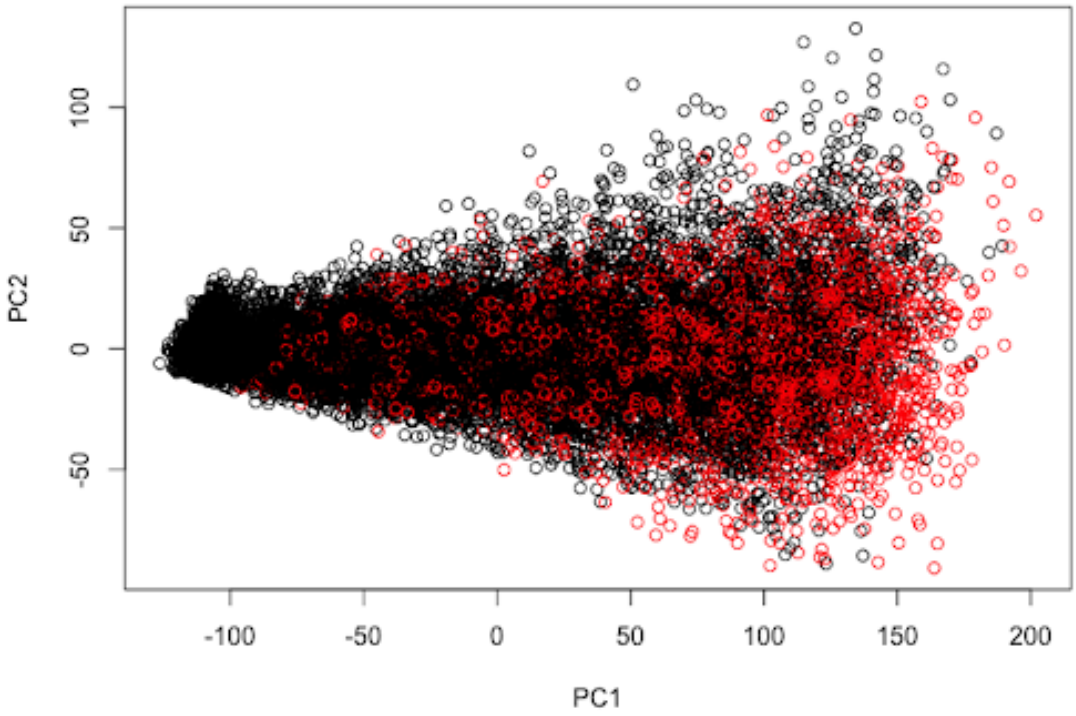
component it captures the highest amount of triples and stolen bases. Once we get to the seventh principal component much of the data is accounted for at this point, but the most significant variable in this principal component is weight. We narrowed our data down to answer our bigger questions for our project. Which player are considered All Stars in terms of hitting and stolen bases. Two ran who different PCA analysis on this to print out some scatter plots to help us identify the All Stars in both stolen bases and hits.

We created this PCA plot to determine stolen bases. In our data we looked at the height, weight, hits, and stolen bases variables to create this plot **Figure 12**. After running the PCA we decided to use 3 out of 4 PCAs. This is because the first PCA contains 99% of the hits variable, the second PCA contains 99% of the weight variable and the third PCA contains 99% of the stolen bases variable in our data set. This scatter plot shows all of the players stolen bases in our dataset. The players that are red are our All-Stars. The PCAs came out as negative but still determines the important variables for stolen bases. Sb[,3] on the y-axis is looking at the relationship between amount of hits and how many stolen bases that player has and sb[,1] on the y-axis looks at the weight of the player. From our results, we can determine that the lighter players who have the most his will in relation have the most stolen bases.



(12)

In our graph (**Figure 13**) we can see that much of or red data points, which are our All Stars, are swayed to the right of our scatter plot. On the x-axis the PC1 is looking at the mostly hits, games, and RBIs. On the y-axis, PC2 specifies mostly strikeouts a player has. So when we interpret this graph we can see that when a player has a lower PC2 score that means that the player has a lower amount of strikeouts and a higher amount of hits. In order to be an All Star, the lower strikeouts that the player has and the higher hits that player has, makes that player an All Star.



(13)

5 Predicting HR, SB, and OPS

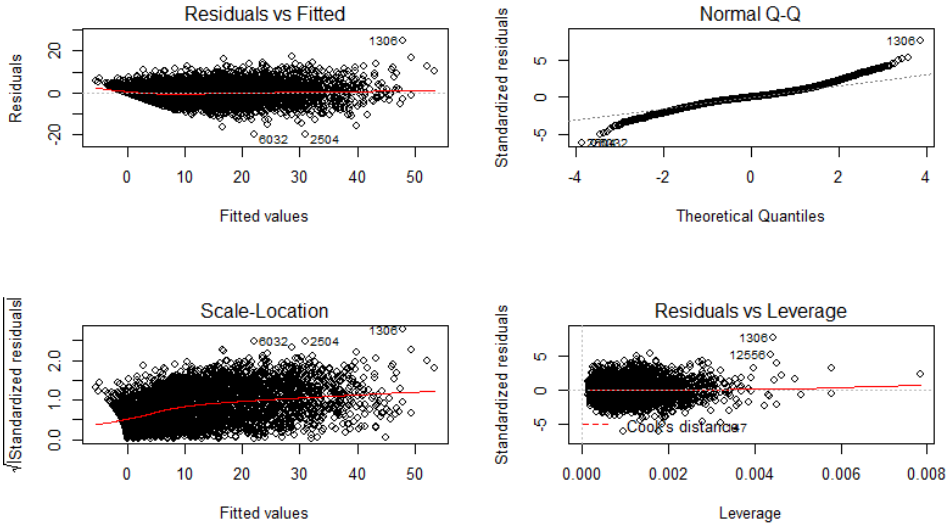
5.1 Linear Regression Models

In regards to the methodologies and techniques used to predict *HR, SB,* and *OPS*, our group decided to use linear regression. In addition to developing linear regression models for each variable, our group used subset selection to understand the appropriate size each of the models and finally we applied shrinkage methods lasso and ridge regression to see if the models generated using best subset selection contained predictors whose coefficients could be shrunk essentially to zero or exactly to zero. However before running subset selection and shrinkage, our group wanted to create simple and multiple linear regression models to see what predictors would be significant in predicting *HR, SB,* and *OPS*.

Now taking a look at our linear model to predict *HR*, our first model we see uses the following predictors: *AB, R, H, Weight, RBI* and *All-Star* are statistically significant. Since all predictors are significant, we reject the Null hypothesis. Moreover, the model's goodness of fit is measured by the adjusted R-squared which is 0.8879 and the model's MSE is 10.7. Below are figures containing the models results. **Figure 15** shows a diagnostic plot of our linear model. From this visual, we see our model captures non-linearity and doesn't have any high leverage points. Moreover, there appears to be slight heteroscedasicity, as seen in the bottom left corner, and the Q-Q plot shows us that our distribution of residuals is skewed.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -38.115705   1.850737 -20.595  < 2e-16 ***
log(AB)       0.073085   0.124997   0.585    0.559
R             0.175144   0.003925  44.620  < 2e-16 ***
H            -0.145106   0.002756 -52.642  < 2e-16 ***
log(weight)   7.096537   0.328981  21.571  < 2e-16 ***
RBI           0.350460   0.002841 123.369  < 2e-16 ***
All.Star1     0.607897   0.106023   5.734 1.02e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.268 on 8826 degrees of freedom
Multiple R-squared:  0.8902,    Adjusted R-squared:  0.8901
F-statistic: 1.192e+04 on 6 and 8826 DF,  p-value: < 2.2e-16
```

MODEL/hr lm summary.png (14)

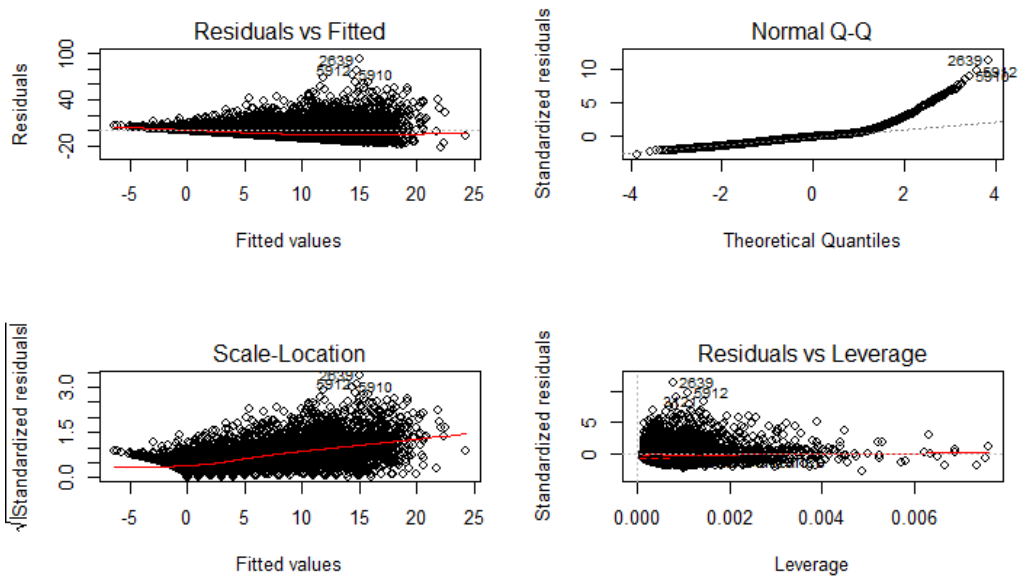


MODEL/hr linear model 11₁1.*png* (15)

Running our linear model for *SB* shows us that *PC1, PC3, H, All.Star*, and the interaction term *PC1*PC3* are statistically significant. Since all predictors are significant, we reject the Null hypothesis. Furthermore, the model's adjusted R-squared is .30 and it has a RSE of 8.07. Below are figures containing the models results. **Figure 17** shows a diagnostic plot of our linear model. From this visual, we see our model captures non-linearity, and doesn't have any high leverage points. The Q-Q plot however shows us that our distribution of residuals is very skewed. Finally, the graphic on the bottom left shows that heteroscedasicity is occurring.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.901e+01  8.532e-01 -22.277  < 2e-16 ***
PC1         -1.908e-01  9.153e-03 -20.845  < 2e-16 ***
PC3          1.140e-01  5.862e-03  19.446  < 2e-16 ***
H            3.119e-01  1.051e-02  29.684  < 2e-16 ***
All.Star    -1.072e+00  2.531e-01  -4.236  2.3e-05 ***
PC1:PC3      7.467e-04  8.170e-05   9.139  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.074 on 8827 degrees of freedom
Multiple R-squared:  0.3089,    Adjusted R-squared:  0.3085
F-statistic:   789 on 5 and 8827 DF,  p-value: < 2.2e-16
```

sb.png (16)

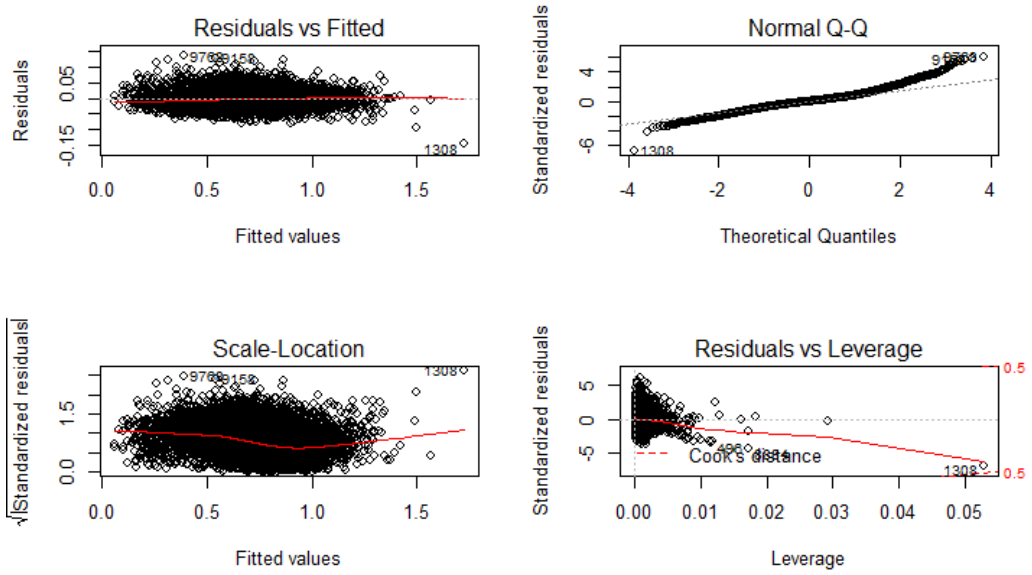


sb dia plot.png (17)

Lasty, we take a look at OPS. Our linear model has the predictors; *AVG, ISO, X3B, RBI,* the polynomial term *BB* and the interaction term *AVG*ISO* are all predictors that are statistically significant. Since all predictors are significant, we reject the Null hypothesis. In addition, the model's goodness of fit is at an adjusted R-squared of 0.986 and the model has an RSE of 0.021. Below are figures containing the models results. **Figure 19** shows a diagnostic plot of our linear model. From this visual, we see our model captures non-linearity, and homoscedasicity. However, just like with our *SB* model, the Q-Q plot shows us that our distribution of residuals are skewed. Moreover, there is a potentially high leverage point.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.374e-02  1.814e-03   13.09   <2e-16 ***
I(BB^2)      8.760e-06  1.364e-07   64.23   <2e-16 ***
AVG          2.107e+00  8.088e-03  260.55   <2e-16 ***
ISO          1.308e+00  1.076e-02  121.55   <2e-16 ***
X3B         -1.084e-03  1.166e-04   -9.30   <2e-16 ***
RBI         -3.542e-04  1.319e-05  -26.84   <2e-16 ***
AVG:ISO     -8.691e-01  4.120e-02  -21.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02199 on 8826 degrees of freedom
Multiple R-squared:  0.986,     Adjusted R-squared:  0.986
F-statistic: 1.036e+05 on 6 and 8826 DF,  p-value: < 2.2e-16
```

ops.png (18)



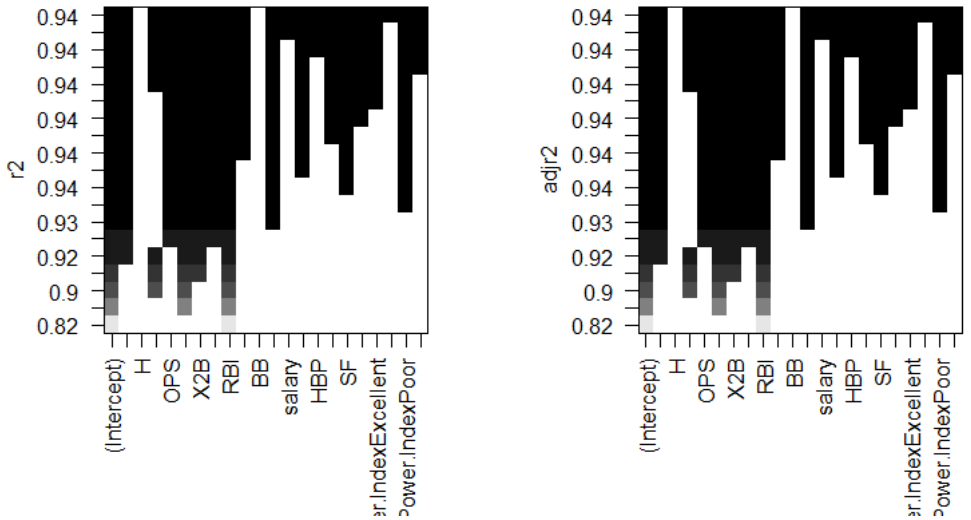
ops dia plot.png (19)

After developing these simple to complex models, our group got a much better understanding of what predictors could be useful in order to predict the variables of interest. Furthermore, our group noticed how the diagnostic plots were showing how our models weren't as well fit as we thought. Therefore, our group used the technique called subset selection in order to develop the best models possible.

5.2 Best Subset Selection

After running simple and multiple linear regression our group was beginning to understand what variables to include in a final model. But in order to determine what the best possible model could be for predicting *HR, SB,* and *OPS*, our group used subset selection. After performing forward, backward, and best subset selection, we discovered that best subset yielded the best models for predicting the outcomes of interest. We checked the size of *M*, the amount of variables included in each model, by using Validation error and k-fold Cross Validation.

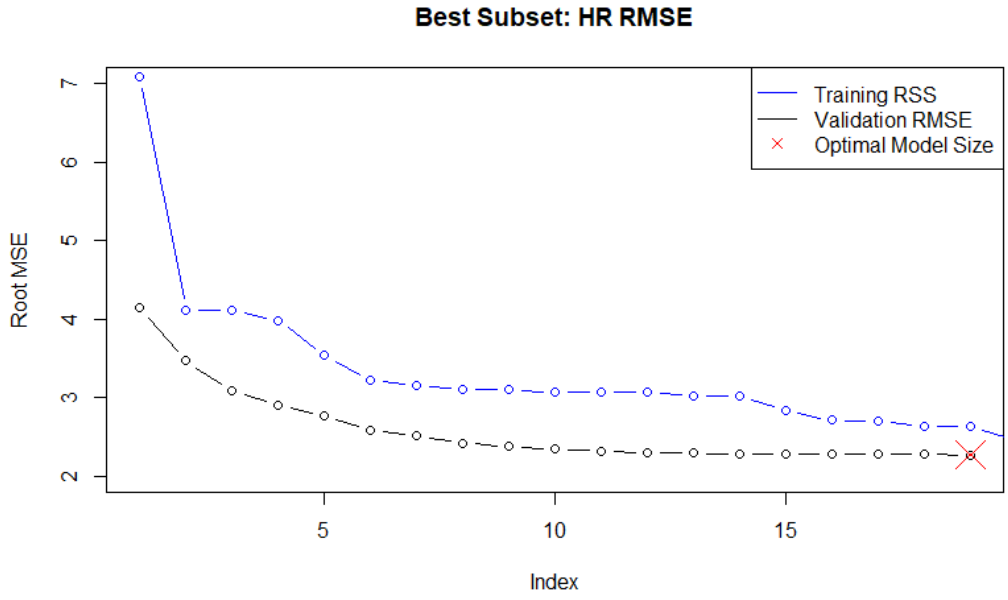
For the HR model, best subset selected a model that included every predictor in our dataset, yielding a model with 19 predictors. The adjusted R-squared for this model is 0.94 and has a RMSE of 2.3. **Figure 20, 21, 22** display the adjusted training error as well as the validation error and cross validation error associated with the different size models. The variables *Runs, Hits, OPS, ISO, Doubles, Triples, RBI, SB, SO, Salary, IBB, HBP, SH, SF, PA, Power.IndexExcellent, Power.IndexGreat, Power.Indexpoor,* and *All.Star1* show the highest R-sqaured and adjusted R-squared along with the lowest RMSE, the two approaches used to estimate our test error is Validation set approach and k-fold approach of the test data. In other words, best subset selection built a full model for predicting *HR*.



r2 adjr2.png

(20)

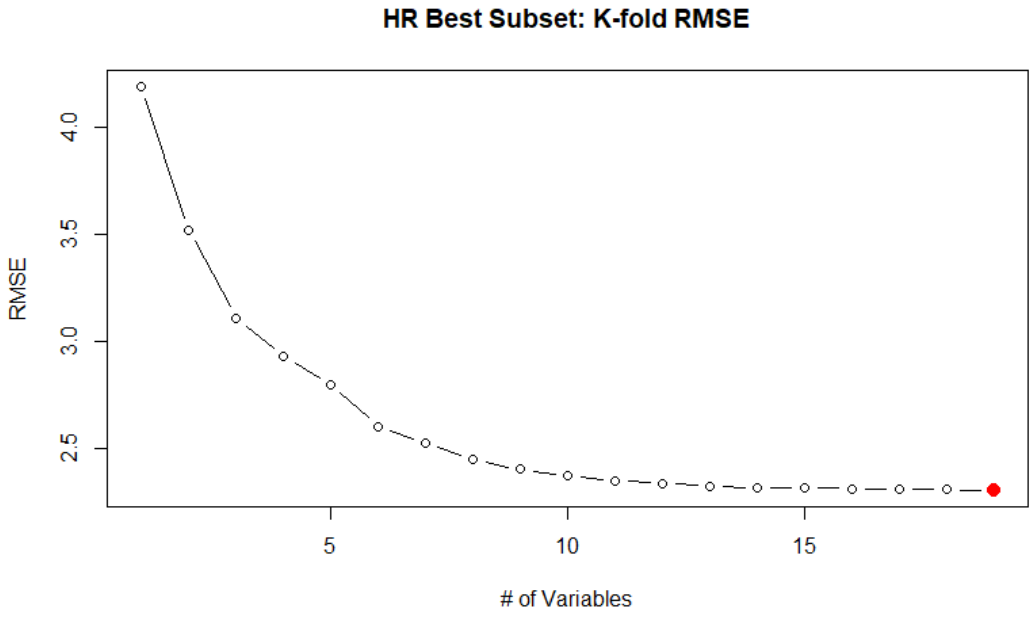
Graph of RMSE validation approach



HR RMSE.png

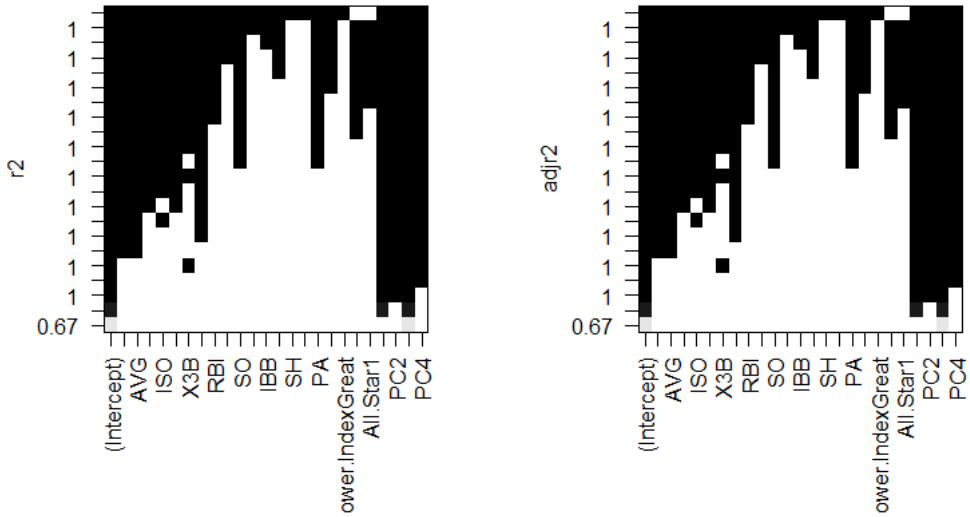
(21)

Graph of RMSE k-fold



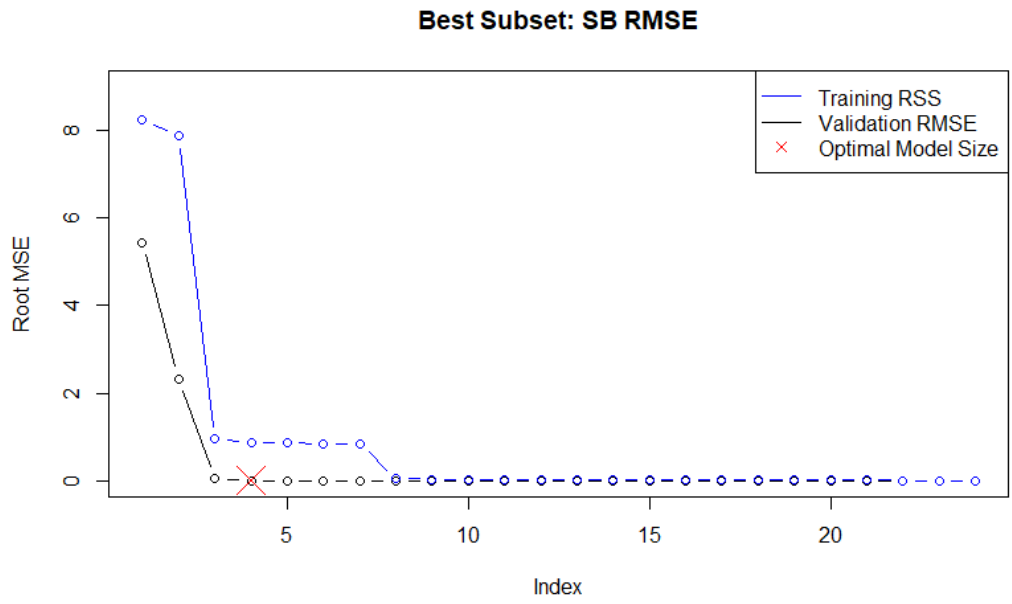
kfold.png (22)

For the SB model, best subset selected a model that included only four predictors. These predictors happen to be the Principal Components; PC1-4, we generated in our Principal Component Analysis. The model that best subset made for predicting SB has an R-squared of .97 and an RMSE of .00063. **Figure 23, 24, 25** display the adjusted training error as well as the validation error and cross validation error associated with the different size models. PC1, PC2, PC3, and PC4 show the R-sqaured and adjusted R-squared along with the RMSE. This RMSE tells us that our model can essentially predict SB with almost no error whatsoever.



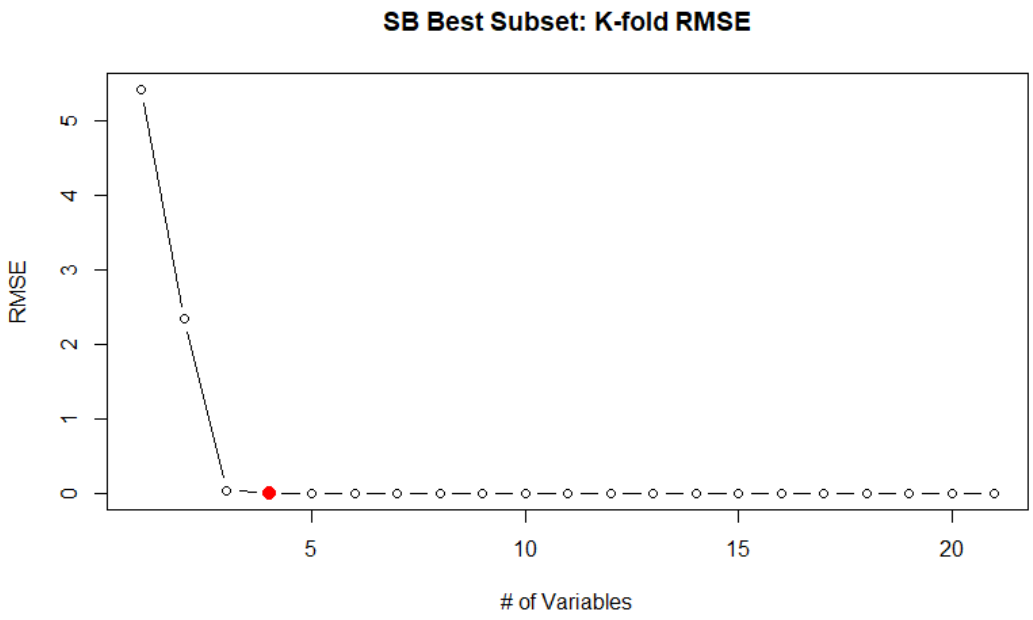
r2 adjr2.png (23)

Graph of RMSE validation approach



SB RMSE.png (24)

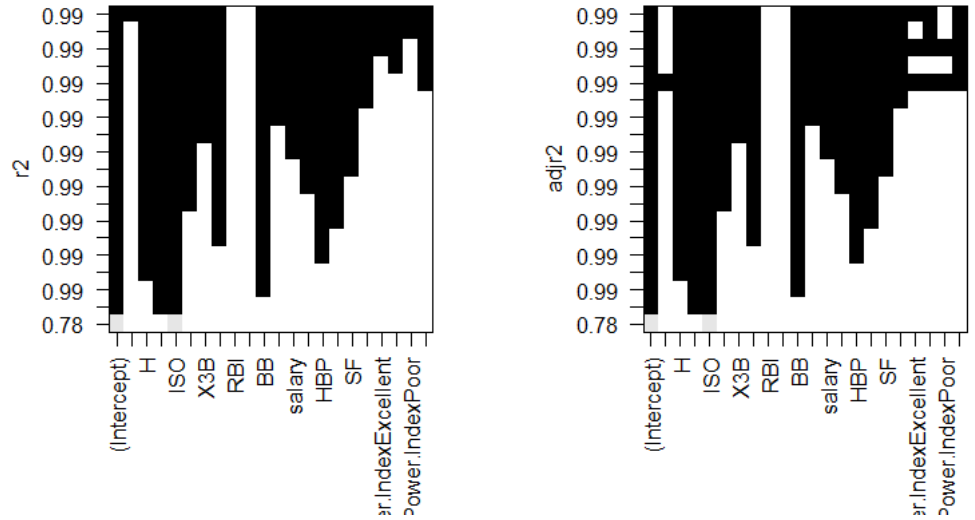
Graph of RMSE k-fold



kfold.png (25)

For the OPS model, best subset selected a model with 17 variables originally. However, after performing two validation metrics; Validation error and k-fold Cross Validation error, we can see in **Figure 28** that k-fold Cross Validation yields a lower test error when we include only 15 variables chosen by best subset instead of 17 variables. **Figure 26, 27, 28** display the adjusted training error as well as the validation error and cross validation error associated with the different size models. Variables *Hits, Average, ISO, Doubles, Triples, HR, BB, SO, Salary, IBB, HBP, SH, SF, PA, Power.IndexGreat, Power.IndexPoor,* and *All.Star1*

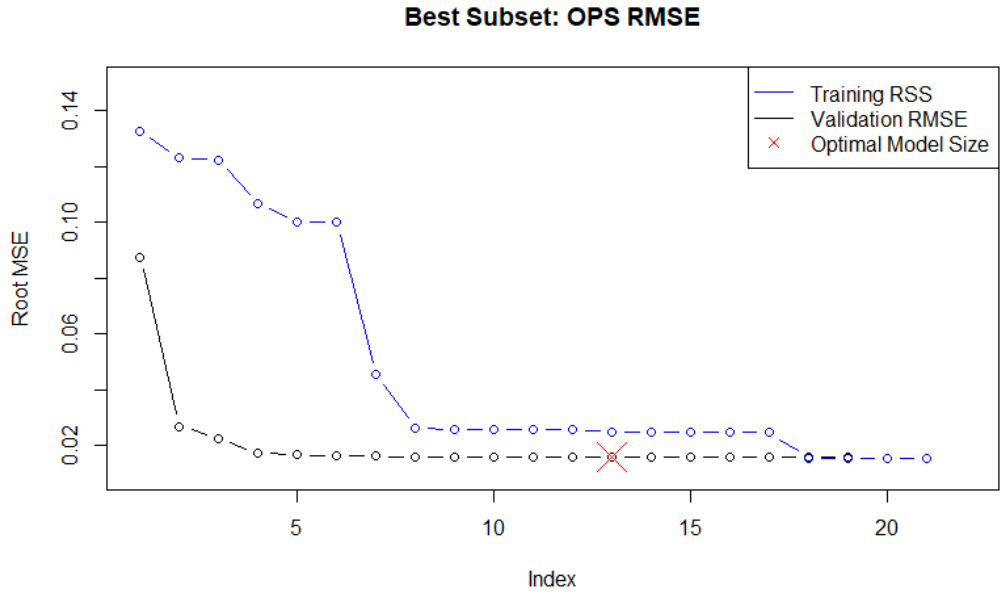
show the highest R-sqaured and adjusted R-squared along with the lowest RMSE for the model. Since our goal in building these regression models is to make the most accurate model using the least amount of predictors, we selected the model when M = 15.



r2 adjr2.png

(26)

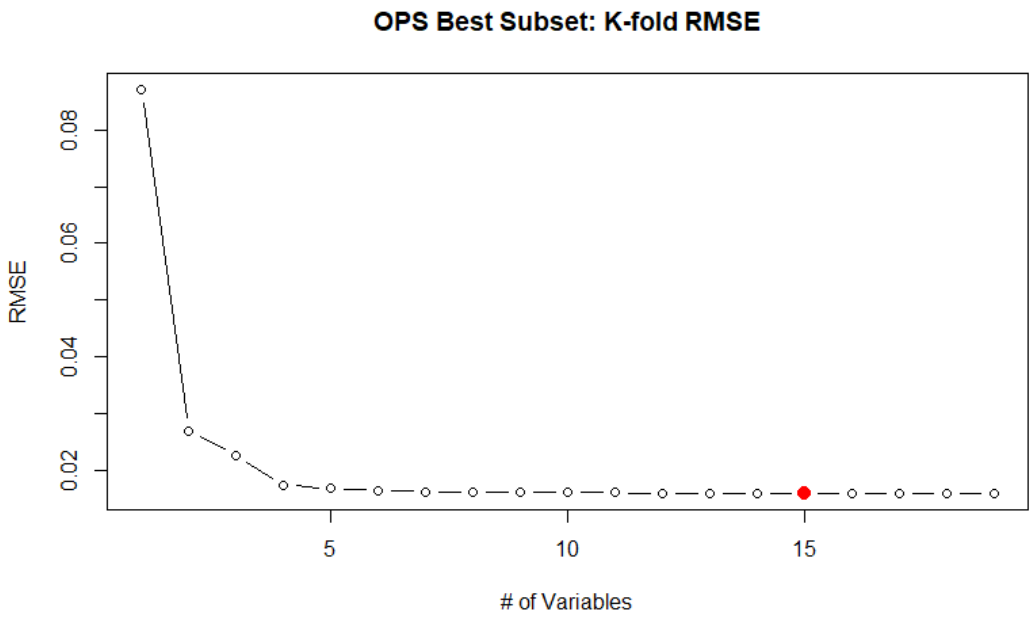
Graph of RMSE validation approach



OPS RMSE.png

(27)

Graph of RMSE k-fold



kfold.png (28)

5.3 Shrinkage Methods

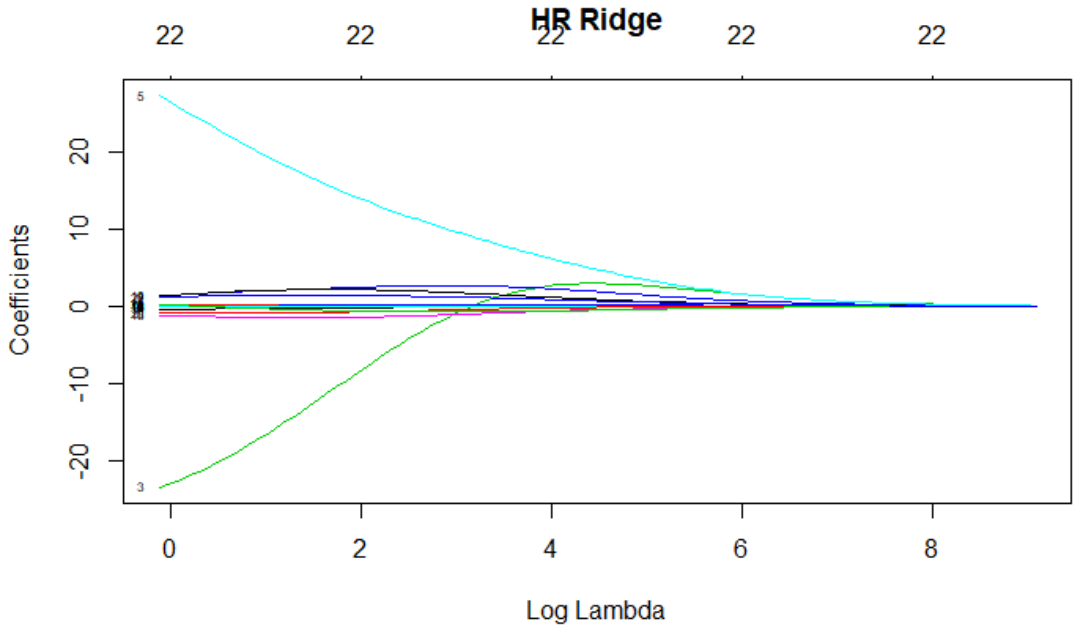
After applying best subset selection to our training data and then applying k-fold cross validation to estimate the test error in picking the optimal size M for each model, our group implemented shrinkage penalties on the models to see how much their coefficients would shrink. We started with Ridge Regression, which is a technique for lower the potentially high variance in our test error. This high variance could appear due to least squares and how the β coefficients our generated.

RIDGE

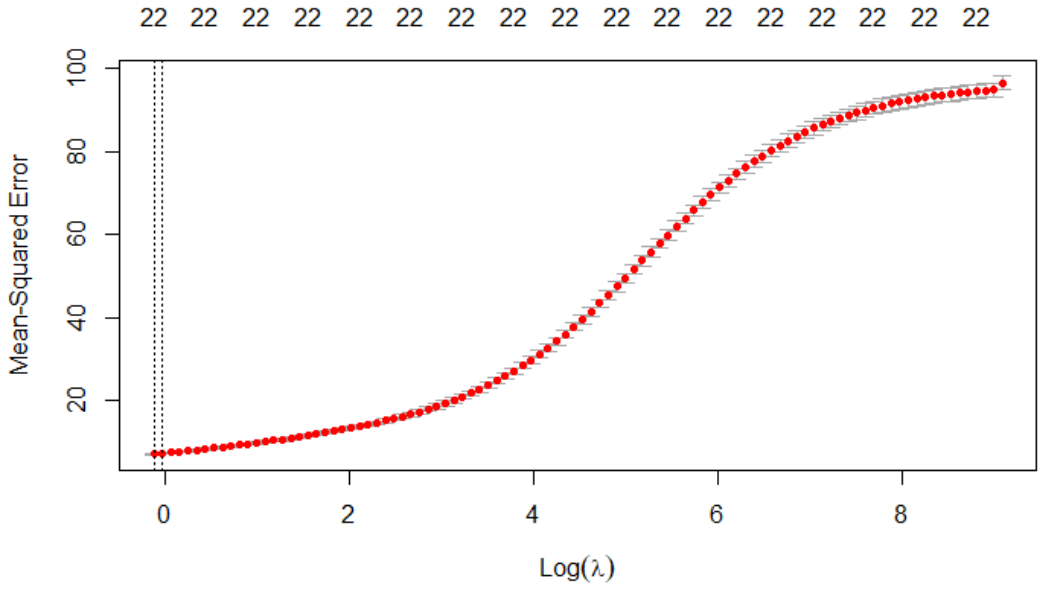
Analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates can have high variation. We have created three Ridge regression models to predict each dependant variable (*HR, SB,* and *OPS*) for our MLB subset. In our Ridge Regression we utilized "Lambda" to tune our parameters, with the goal of lowering the potentially high variance in our test error.

HOME RUNS

Ridge regression for *HR* showed that 2 coefficient estimates for the variables *AVG* and *ISO* were the least significant in predicting the outcome, therefore there coefficient estimates were shrunk the most. The optimal λ was represented as 0.976. **Figure 29** displays what coefficients are shrinking towards zero as λ increases and **Figure 30**.



(29)

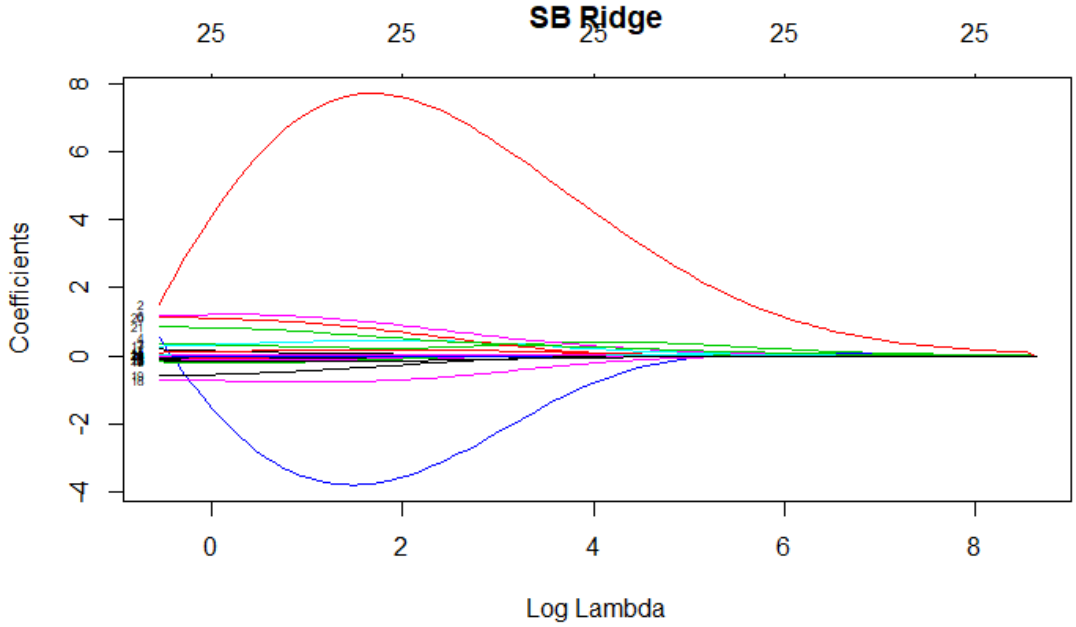


1.png

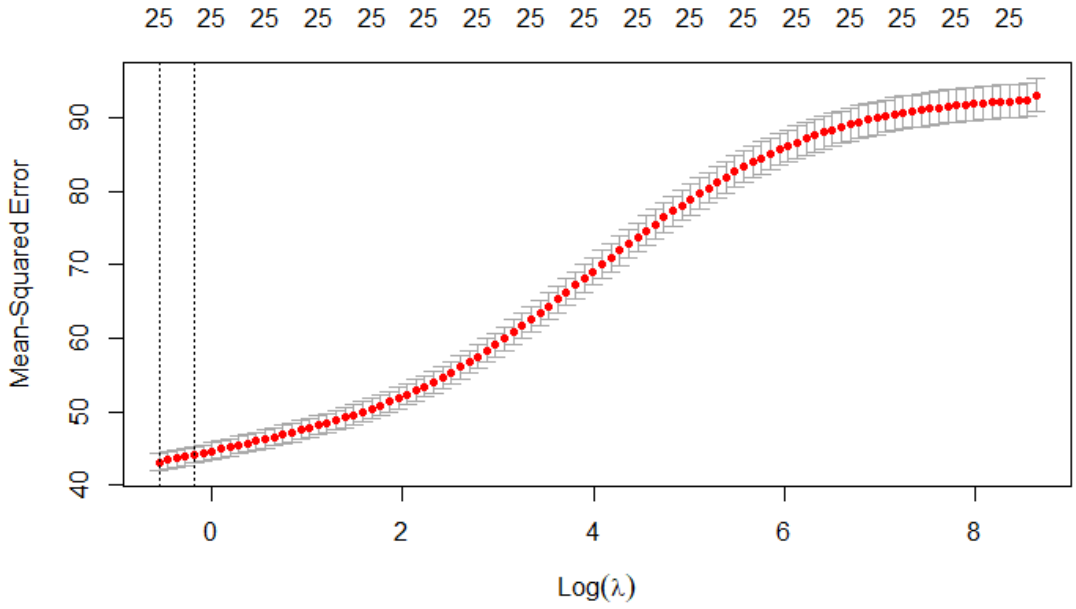
(30)

STOLEN BASE

Ridge regression for *SB* showed that 2 coefficient estimates for the variables *AVG* and *ISO* were the least significant in predicting the outcome, therefore there coefficient estimates were shrunk the most. The optimal λ was represented as 0.126. **Figure 31** displays what coefficients are shrinking towards zero as λ increases and **Figure 32**.



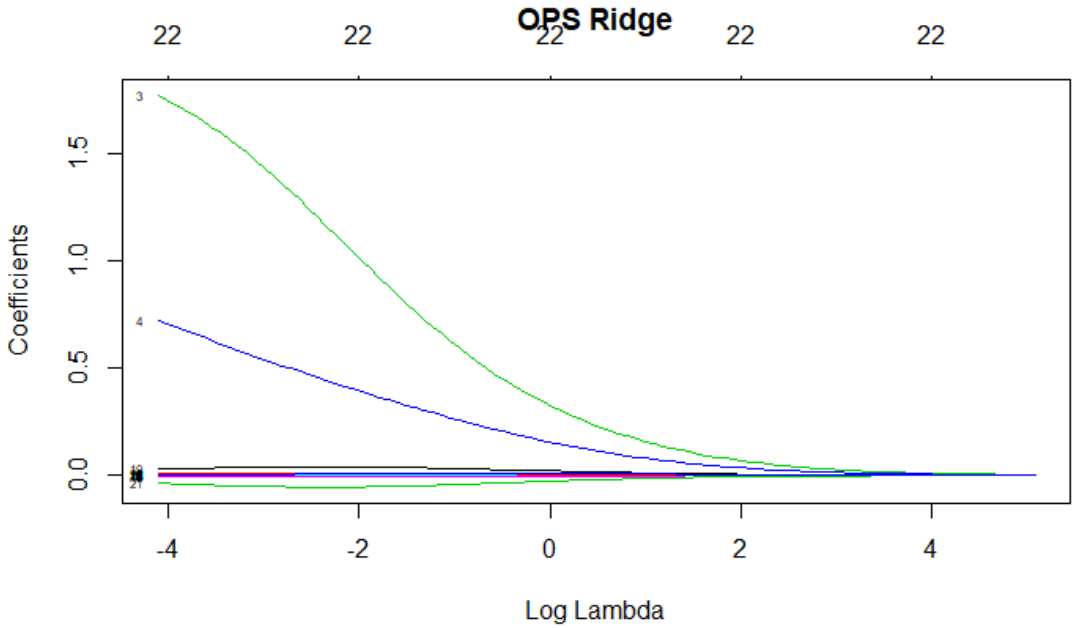
(31)



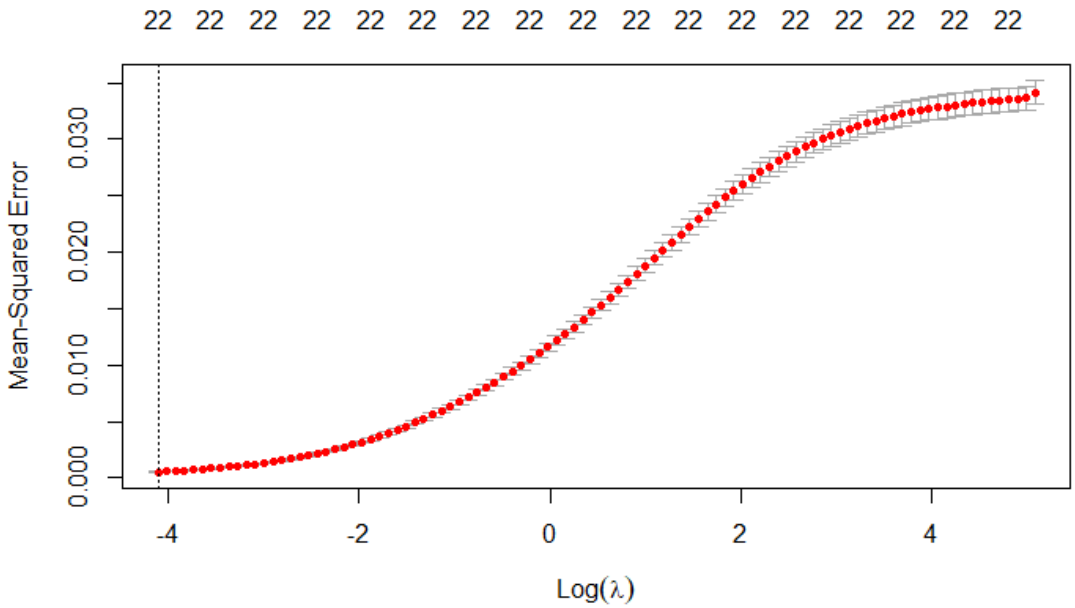
(32)

ON-BASE PLUS SLUGGING

Ridge regression for *OPS* showed that 2 coefficient estimates for the variables *AVG* and *ISO* were the least significant in predicting the outcome, therefore there coefficient estimates were shrunk the most. The optimal λ was represented as 0.00056. **Figure 33** displays what coefficients are shrinking towards zero as λ increases and **Figure 34**.


OPS Ridge
22 22 22 22 22
Coefficients
0.0 0.5 1.0 1.5
3
4
-4 -2 0 2 4
Log Lambda

(33)


22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22
Mean-Squared Error
0.000 0.010 0.020 0.030
-4 -2 0 2 4
Log(λ)

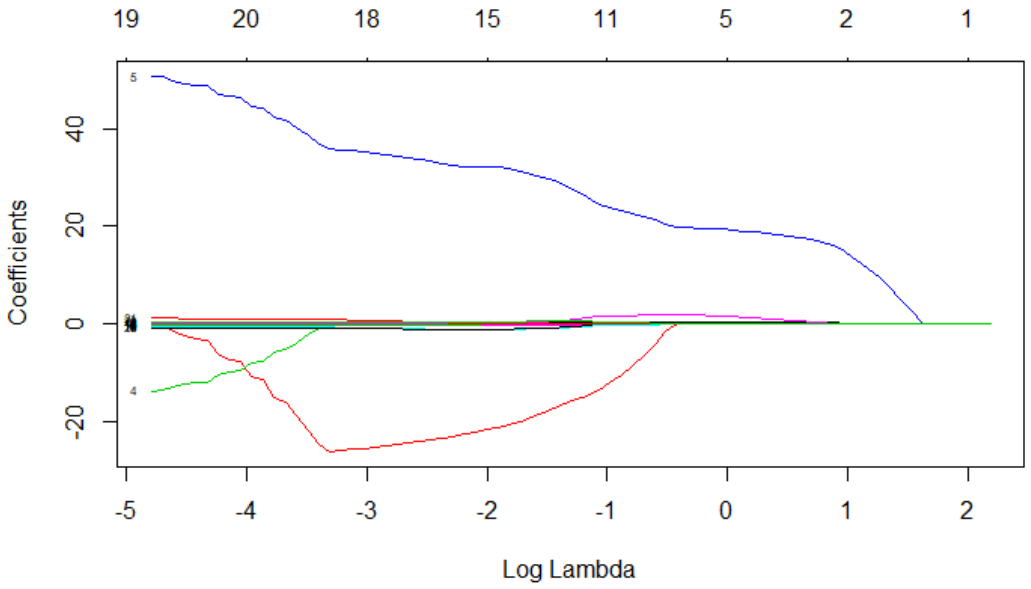
(34)

LASSO

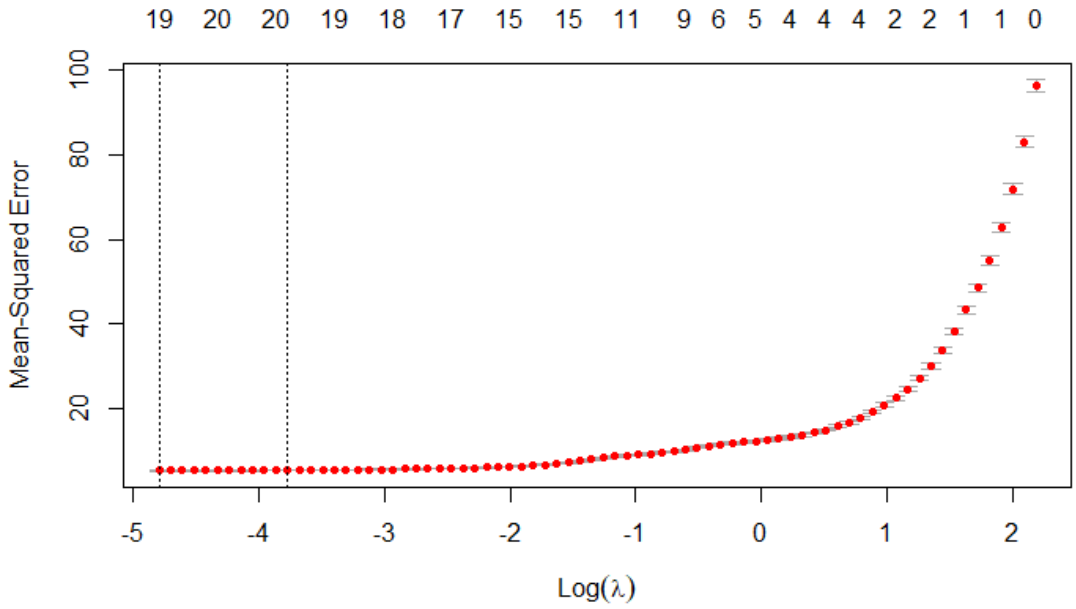
Lasso performs variable selection unlike Ridge regression and therefore some of our variables may be shrunk to zero. With our variables coefficient estimates potentially becoming zero, our group see more precisely of what predictor coefficients are actually relevant in predicting our desired outcomes. Furthermore, we utilize the same models that were in used in Ridge regression.

HOME RUNS

Lasso regression for *HR* showed that 2 coefficient estimates for the variables *OPS* and *ISO* were the least significant in predicting the outcome, therefore there coefficient estimates were shrunk to zero. The optimal λ was represented as 0.0278. There is nothing to be gained by using the independent variables *H* and the dummy variable *Power.IndexExcellent* to predict the outcome of *HR*. **Figure 35** displays what coefficients are shrinking towards zero as λ increases and **Figure 36**.



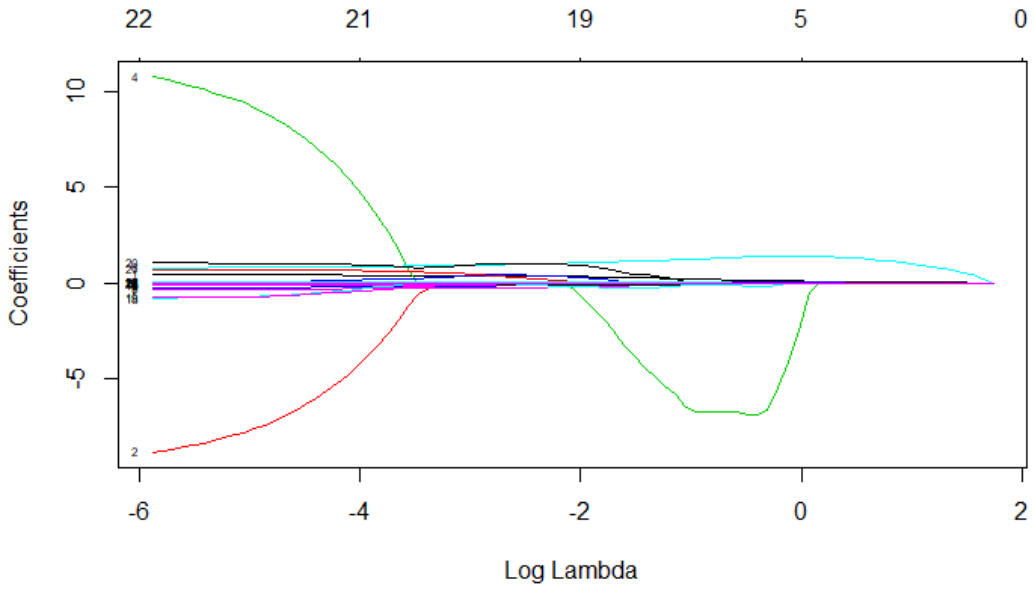
HR.png (35)



(36)

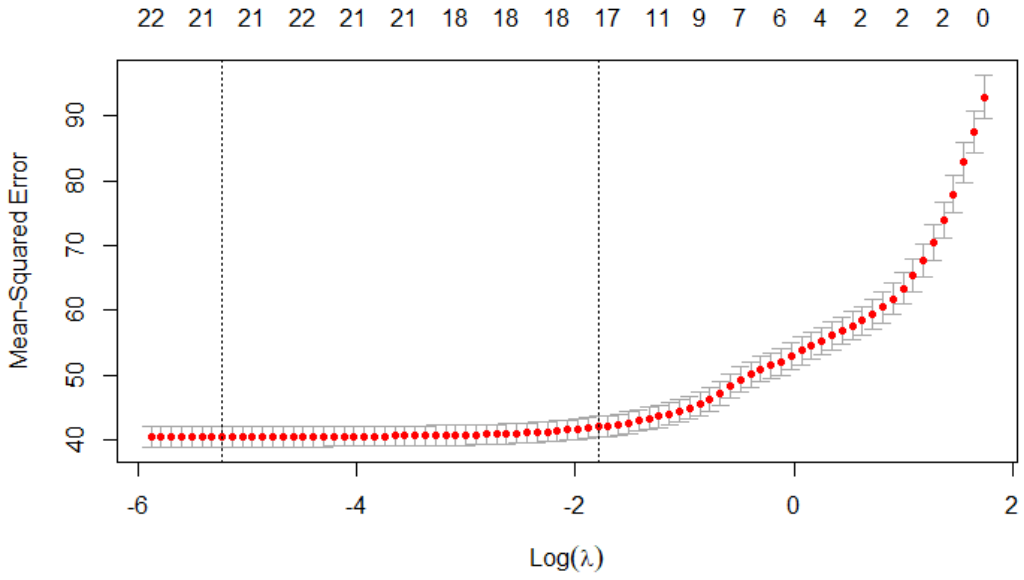
STOLEN BASE

Lasso regression for *SB* showed that 2 coefficient estimates for the variables *AVG* and *ISO* were the least significant in predicting the outcome, therefore there coefficient estimates were shrunk to zero. The optimal λ was represented as 0.126. There is nothing to be gained by using the independent variables *PC1, Power.IndexGreat, PA, SO, OPS* and *AVG* to predict the outcome of *SB*. **Figure 37** displays what coefficients are shrinking towards zero as λ increases and **Figure 38**.



SB.png

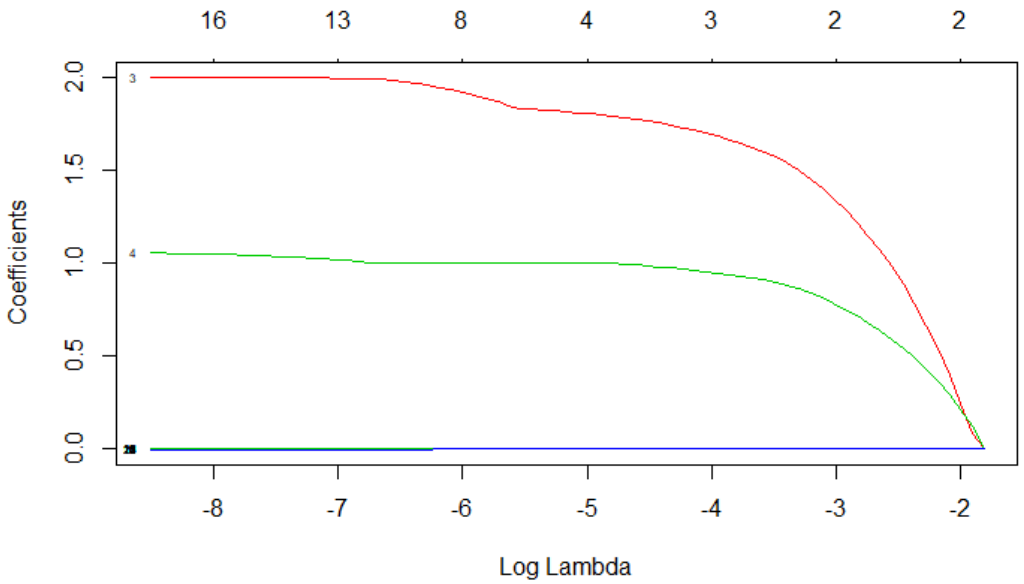
(37)



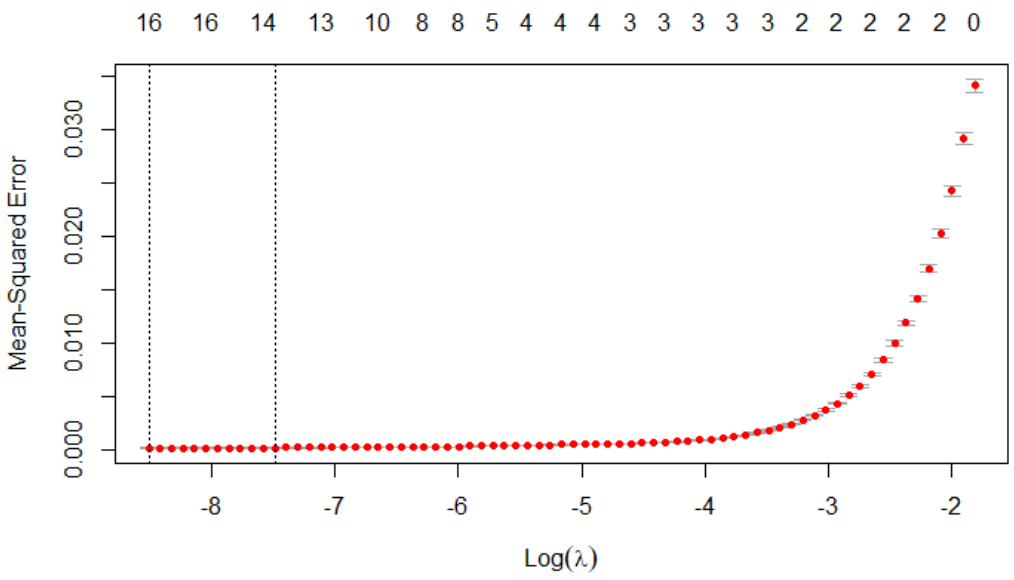
SB 1.png (38)

ON-BASE PLUS SLUGGING

Lasso regression for *OPS* showed that 2 coefficient estimates for the variables *AVG* and *ISO* were the least significant in predicting the outcome, therefore there coefficient estimates were shrunk to zero. The optimal λ was represented as 0.00056. There is nothing to be gained by using the independent variables *R, SB, PA, All.Star1, X3B, Power.IndexExcellent, Power.IndexAverage,* and *Power.IndexGreat* to predict the outcome of *OPS*. **Figure 39** displays what coefficients are shrinking towards zero as λ increases and **Figure 40**.



OPS.png (39)



OPS 1.png (40)

6 All-Star Prediction

6.1 Logistic Model

For the logistic model, we used the predictors: *OPS, Salary, RBI, SO, R, IBB,* and *SB* and the interaction term *RBI*R*, to predict whether a player is going to be an All Star or not. **Figure 41** contains the summary of this model. All of the predictors in our model reject the Null Hypothesis, meaning all of them are significant in predicting All-Star. The interaction term's, *RBI*R*, main effects are also significant in the classification of All-Stars. The coefficient estimates for the predictor *OPS* is .2218, meaning a .001 unit increase results in the log odds increasing our prediction by .22.

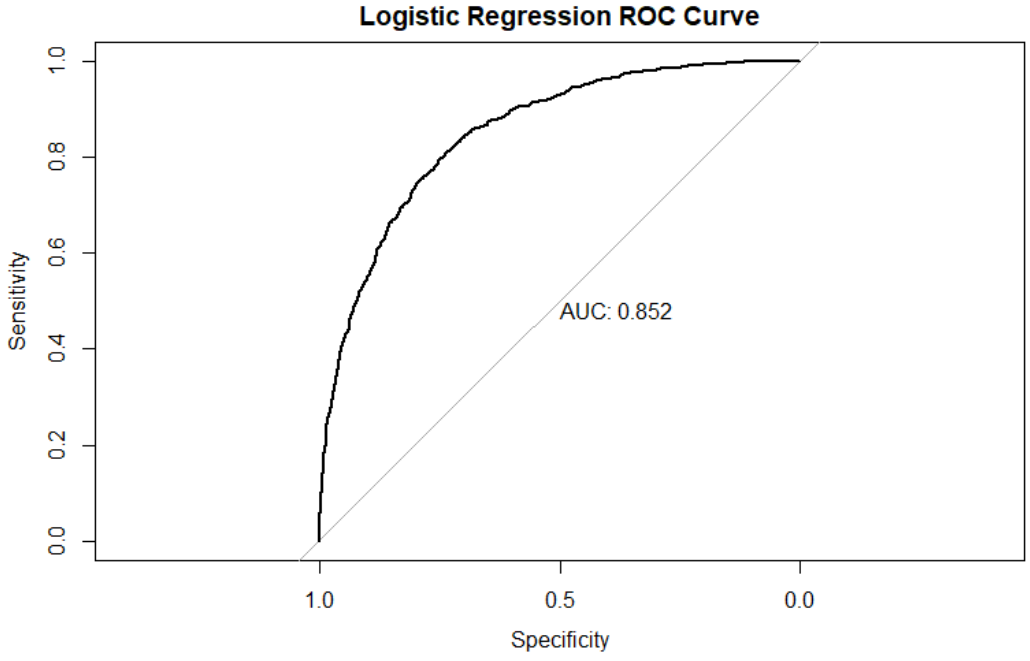
```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.792e+00  2.351e-01 -20.386  < 2e-16 ***
salary       1.346e-07  7.665e-09  17.563  < 2e-16 ***
SO          -1.560e-02  1.320e-03 -11.821  < 2e-16 ***
RBI          3.671e-02  3.566e-03  10.294  < 2e-16 ***
R            2.392e-02  3.197e-03   7.483 7.24e-14 ***
IBB          6.110e-02  9.396e-03   6.503 7.88e-11 ***
OPS          2.218e+00  2.958e-01   7.497 6.54e-14 ***
SB           7.559e-03  3.818e-03   1.980   0.0477 *
RBI:R       -1.701e-04  3.676e-05  -4.626 3.73e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9613.6  on 7630  degrees of freedom
Residual deviance: 6490.9  on 7622  degrees of freedom
AIC: 6508.9

Number of Fisher Scoring iterations: 5
```

log sum.png (41)

The predictive performance of our model can be seen in **Figure 42** which showcases a confusion matrix. The threshold used as our cutoff was .2, meaning we label a player an All-Star if our model predicted his posterior probability to be greater than or equal to .2. Based on our testing data, which is a little over 5,000 observations, we can see that our model has a prediction accuracy of 69.6%. The metric our group is interested in is called Specificity, which measures the percentage of accurate we can classify 1's with our model (the 1's denote All-Stars). The model used correctly classifies Class 0 (Non-All Stars) with 64% accuracy and correctly classifies Class 1 (All Stars) with 88% accuracy. Although our model's misclassification rate is around 30%, our main goal is identifying All-Star's correctly. To better visual the true and false positive rates, **Figure 43** contains a ROC curve of our model which has an AUC of .852.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1922  106
         1 1081  795

               Accuracy : 0.696
                 95% CI : (0.6813, 0.7104)
    No Information Rate : 0.7692
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3789

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6400
            Specificity : 0.8824
         Pos Pred Value : 0.9477
         Neg Pred Value : 0.4238
             Prevalence : 0.7692
         Detection Rate : 0.4923
   Detection Prevalence : 0.5195
      Balanced Accuracy : 0.7612

       'Positive' Class : 0
```

log confusion.png (42)



roc.png (43)

6.2 Random Forest

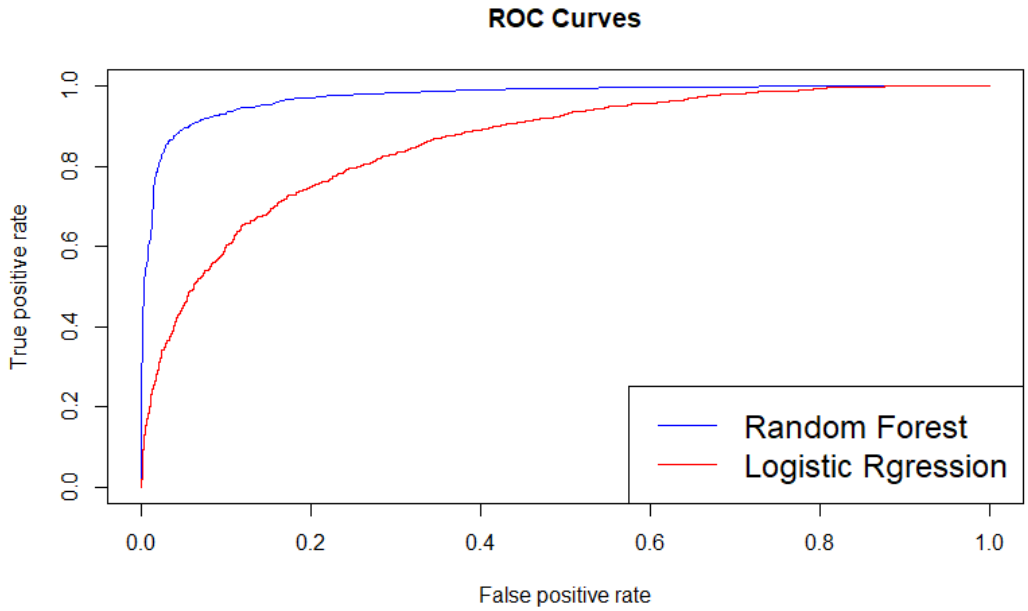
Another methodology we used in predicting if a player will be an All-Star is a technique called Random Forest. When performing random forest we set the maximum amount of predictors to consider per tree to 4, out of the possible 19, and had our number of trees set to 400. Upon performing Random Forest, our group was able to determine the most important predictors were in predicting All-Star; such as variables *salary,*

RBI, and *OPS*. **Figure 44** displays the confusion matrix of our Random Forest, which yields a specificity of 88%. Our group noticed that the Random Forest technique only has a misclassification error rate of 5%! This is a massive difference compared to the Logistic Regression model. Yet, both techniques yield the same predictive accuracy.

For a more in depth look at the comparison on how the two models are performing, **Figure 45** showcases the comparison of our Random Forest ROC curve to our Logistic Regression ROC curve. Based on this plot, we see that Random Forest is clearly the more superior method, since its ROC curve is near optimal and has a much larger AUC than Logistic Regression, in predicting All-Stars.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2863  107
         1  122  786

               Accuracy : 0.9409
                 95% CI : (0.9331, 0.9482)
    No Information Rate : 0.7697
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8344

 Mcnemar's Test P-Value : 0.3549

            Sensitivity : 0.9591
            Specificity : 0.8802
         Pos Pred Value : 0.9640
         Neg Pred Value : 0.8656
             Prevalence : 0.7697
         Detection Rate : 0.7383
   Detection Prevalence : 0.7659
      Balanced Accuracy : 0.9197

       'Positive' Class : 0
```

confusion.png (44)



curves.png (45)

7 Final Results

To conclude, we were able to accurately predict All-Stars in HR, Stolen Bases, and OPS. The purpose of our model is not only to detect All-Star players but to ensure that they are in fact All-Stars. In order to predict an All-Star the variables that were most important to our model are Salary, RBI, and OPS. With these variables we were able to answer our questions. As a result, the final All-Star classification was extremely accurate with our logistic model and our random forest model. We were able to have predicted All-Stars with 88% accuracy when using our logistic model. When we ran our model using random forest we were able to improve on our model by 3% accuracy. Using random forest we are able to predict an All-Star with 91% accuracy, which is the best result for our model.