REGRESSION ANALYSIS OF BASEBALL STATISTICS

ISyE 6414 Regression Analysis Project Final Report

Prepared by: Felipe Imanishi, Wonhee Lee, Sagar Babu, Team number: 22

Data source: https://www.kaggle.com/seanlahman/the-history-of-baseball/data

TABLE OF CONTENTS

INTRODUCTION

Baseball is one of the most popular sports in the United States of America. Despite being a team game, the flow of the game is such that it has natural breaks and individual performance often decides the outcome of games, this leads to easy record-keeping and statistics tracking of individuals and teams. Our dataset tracks the statistics of every aspect of the game all the way back to 1871 giving us almost 150 years' worth of rich data to work with.

Reason for study

One of the major areas where analytics has provided value is sports. At the forefront of sports analytics is baseball analytics, well documented by the "Moneyball" revolution. A unique aspect of baseball is that it is rich with statistics like no other sport; the game naturally involves stops between plays which allows teams to record a bevy of statistics. With such information, teams have attempted to gain a competitive edge by developing models that help them understand how to maximize productivity on the field. Along these lines, our analysis focuses on discovering models that explain which factors contribute to scoring and preventing runs, as well as which attributes are instrumental in getting inducted to the hall of fame.

Expectation from study

We expect the models we develop to have reasonable explanatory power and to generate coefficients that help us understand which factors contribute at which magnitudes to our respective response variables.

In particular for the batting analysis we expect home runs to have the largest impact on runs scored which is the response. For the pitching analysis we expect preventing home runs to to have the largest impact on reducing earned run average which is the response. For the hall of fame analysis, we expect either number of hits or home runs to contribute the most to getting selected.

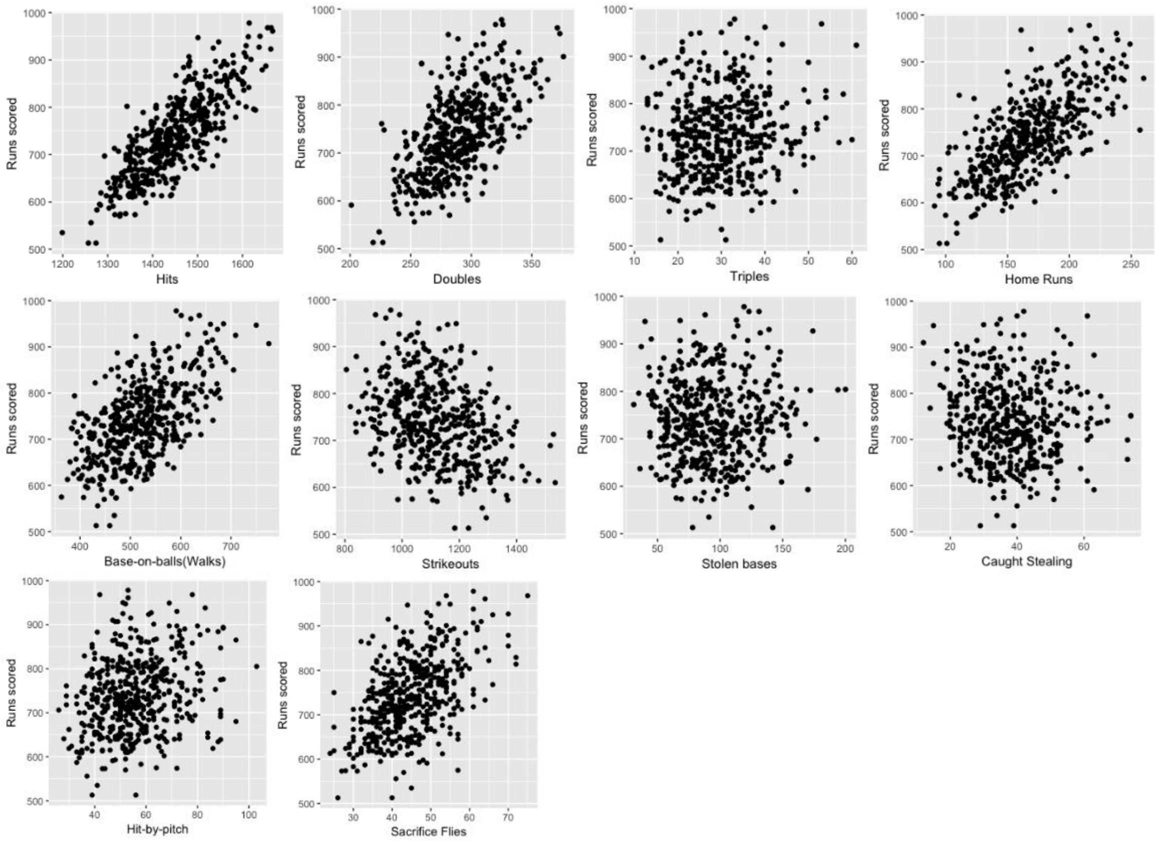
Raw data

The data has baseball statistics from 1871 to 2015 with more than 2800 data points and 48 features. We downloaded this dataset in csv format from Kaggle.

TEAM OFFENSE

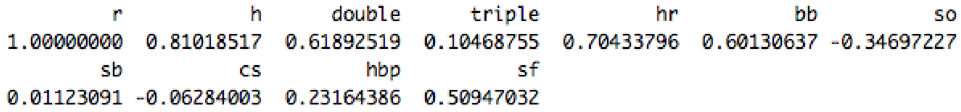
The focus of this section is to see which offensive categories contribute the most to the ultimate objective of scoring runs. In the team offense data in the database, we used hits, doubles, triples, home runs, walks, strikeouts, stolen bases, base-on-balls, strikeouts, stolen bases, caught stealing, hit-by-pitch, and sacrifice flies as predictors and the number of runs scored as our response to build a multiple linear regression model.

1. Exploratory Data Analysis (Scatter plots)



- We can observe a linear relationship between many of the categories and runs scored.
- Strikeouts are negatively associated as expected; the more strikeouts a team has, the less likely it is for the team to score runs.
- Stolen bases and caught stealing are not as strongly correlated; this is expected as they do not affect run-scoring as directly as the other categories do.

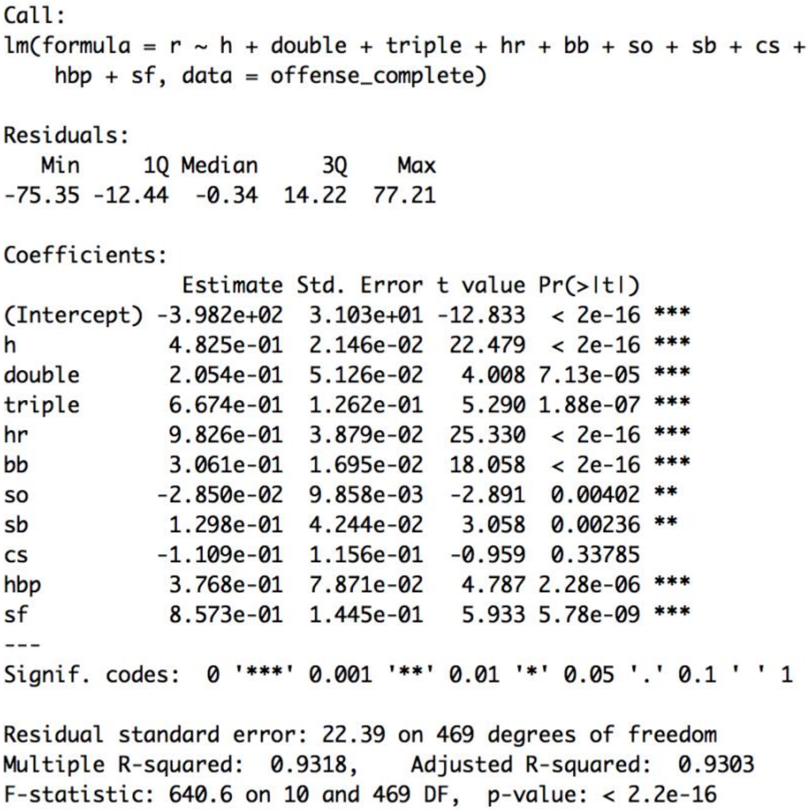
2. Exploratory Data Analysis (Correlation)



```
          r           h      double      triple          hr          bb          so 
 1.00000000  0.81018517  0.61892519  0.10468755  0.70433796  0.60130637 -0.34697227 
         sb          cs         hbp          sf 
 0.01123091 -0.06284003  0.23164386  0.50947032 
```

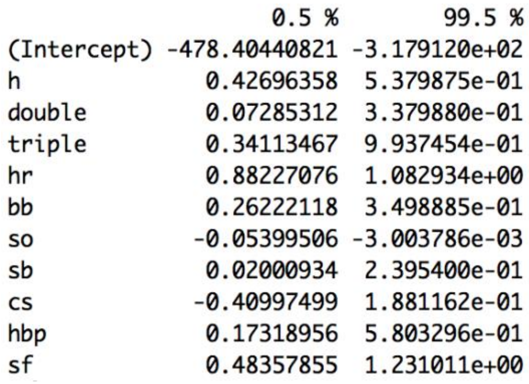
3. Multiple Linear Regression Model

Below is a summary of the trained model.



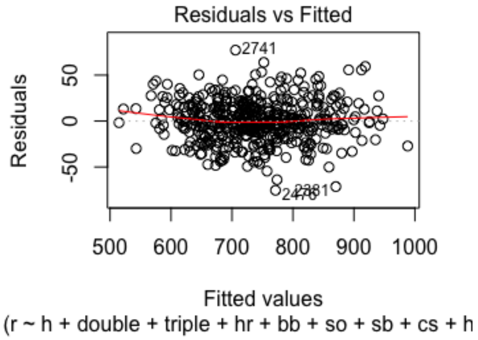
```
Call:
lm(formula = r ~ h + double + triple + hr + bb + so + sb + cs + 
    hbp + sf, data = offense_complete)

Residuals:
   Min     1Q Median     3Q    Max 
-75.35 -12.44  -0.34  14.22  77.21 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.982e+02  3.103e+01 -12.833  < 2e-16 ***
h            4.825e-01  2.146e-02  22.479  < 2e-16 ***
double       2.054e-01  5.126e-02   4.008 7.13e-05 ***
triple       6.674e-01  1.262e-01   5.290 1.88e-07 ***
hr           9.826e-01  3.879e-02  25.330  < 2e-16 ***
bb           3.061e-01  1.695e-02  18.058  < 2e-16 ***
so          -2.850e-02  9.858e-03  -2.891  0.00402 ** 
sb           1.298e-01  4.244e-02   3.058  0.00236 ** 
cs          -1.109e-01  1.156e-01  -0.959  0.33785    
hbp          3.768e-01  7.871e-02   4.787 2.28e-06 ***
sf           8.573e-01  1.445e-01   5.933 5.78e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.39 on 469 degrees of freedom
Multiple R-squared:  0.9318,	Adjusted R-squared:  0.9303 
F-statistic: 640.6 on 10 and 469 DF,  p-value: < 2.2e-16
```

- As we can see, the predicting variables with the highest coefficients are hits, triples, home runs, and sacrifice flies.
- Home runs and sacrifice flies have a particularly high coefficient because these categories, by definition, mean that at least one run is scored. (could be more for HRs)
- We left the data unscaled here for interpretation purposes: this way, the coefficient clearly implies how many runs are scored per one unit increase in each variable.
- It also makes sense that triples' coefficient is high, because it only requires a hit or a sacrifice fly, or in some cases a simple ground ball to score a run afterwards.
- A strong R-squared value of 0.9318 indicates that the explanatory power of this model is high.
- All predictors except caught stealing are significant at alpha = 0.01.
- All confidence intervals at the 99% confidence level below except for caught stealing do not intersect 0, illustrating this again.

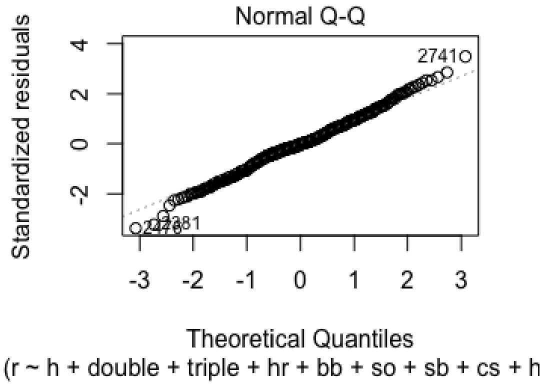


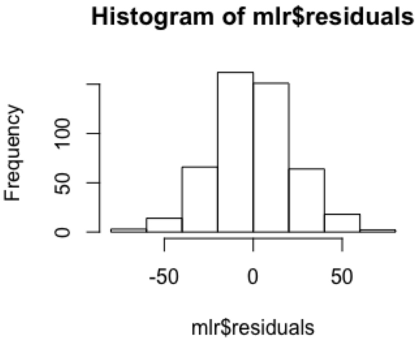
```
                    0.5 %        99.5 %
(Intercept) -478.40440821 -3.179120e+02
h              0.42696358  5.379875e-01
double         0.07285312  3.379880e-01
triple         0.34113467  9.937454e-01
hr             0.88227076  1.082934e+00
bb             0.26222118  3.498885e-01
so            -0.05399506 -3.003786e-03
sb             0.02000934  2.395400e-01
cs            -0.40997499  1.881162e-01
hbp            0.17318956  5.803296e-01
sf             0.48357855  1.231011e+00
```

4. Goodness of fit analysis

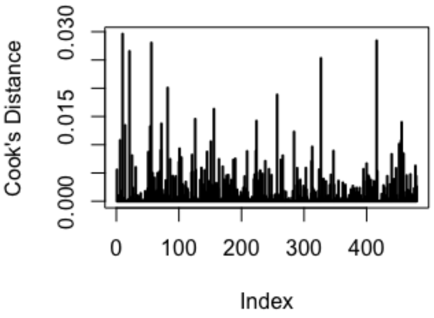


- The residuals vs. fitted values show residuals centered around zero without any particular pattern.
- The variance of residuals appear constant across fitted values.



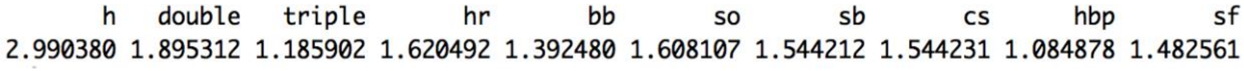


- The Normal Q-Q plot shows a straight line for the most part, demonstrating that the normality assumption holds.
- The histogram of residuals confirms the above.
- However, there is deviation on the edges, particularly for the lowest values. This tells us that normality may not hold up for very low values of runs scored.



- There appear to be no outliers according to Cook’s distances measured.

5. Multicollinearity



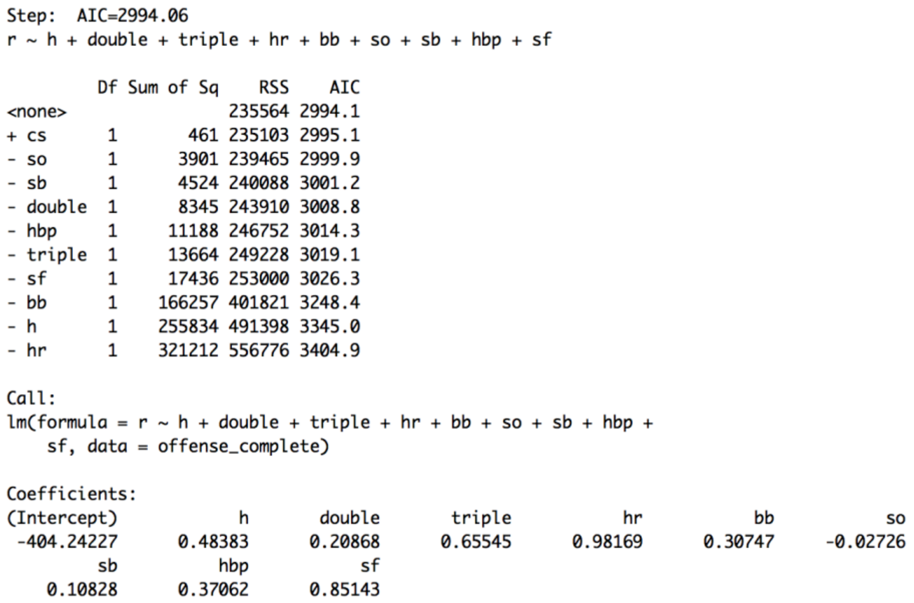
```
       h   double   triple       hr       bb       so       sb       cs      hbp       sf 
2.990380 1.895312 1.185902 1.620492 1.392480 1.608107 1.544212 1.544231 1.084878 1.482561 
```

- None of the VIF values above are greater than 10; thus, we conclude that there is no multicollinearity between the predictors.

6. Variable selection

While 10 predictors is not too many, some organizations may prefer a simpler model, or a model that optimizes the bias-variance tradeoff. In this situation, we can explore 2 options, using AIC and lasso regression for variable selection.

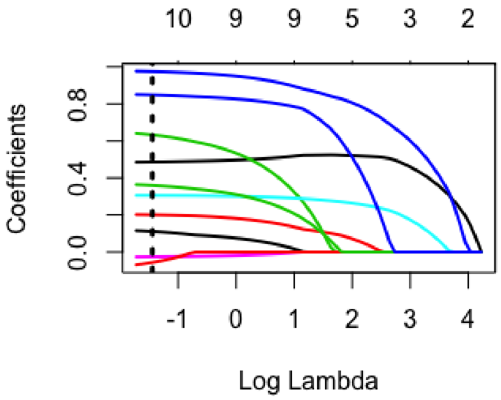
a) AIC



```
Step:  AIC=2994.06
r ~ h + double + triple + hr + bb + so + sb + hbp + sf

         Df Sum of Sq    RSS    AIC
<none>                235564 2994.1
+ cs      1       461 235103 2995.1
- so      1      3901 239465 2999.9
- sb      1      4524 240088 3001.2
- double  1      8345 243910 3008.8
- hbp     1     11188 246752 3014.3
- triple  1     13664 249228 3019.1
- sf      1     17436 253000 3026.3
- bb      1    166257 401821 3248.4
- h       1    255834 491398 3345.0
- hr      1    321212 556776 3404.9

Call:
lm(formula = r ~ h + double + triple + hr + bb + so + sb + hbp +
    sf, data = offense_complete)

Coefficients:
(Intercept)            h       double       triple           hr           bb           so
 -404.24227      0.48383      0.20868      0.65545      0.98169      0.30747     -0.02726
         sb          hbp           sf
    0.10828      0.37062      0.85143
```

- The step function eliminated caught stealing, but kept the other predictors

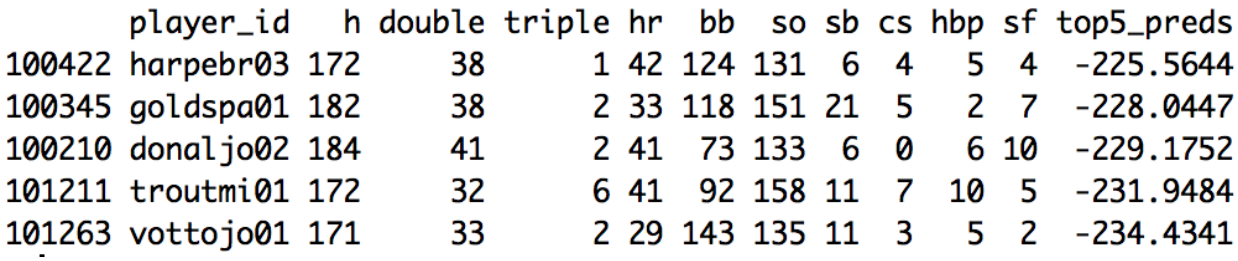
b) Lasso regression (w/ 10-fold cross-validation)



- The lasso regression did not eliminate any predictors in this case.
- The black dotted line illustrates the log of the lambda value computed via the 10-fold cross validation process; we can see that the all predictors remain used at that value.

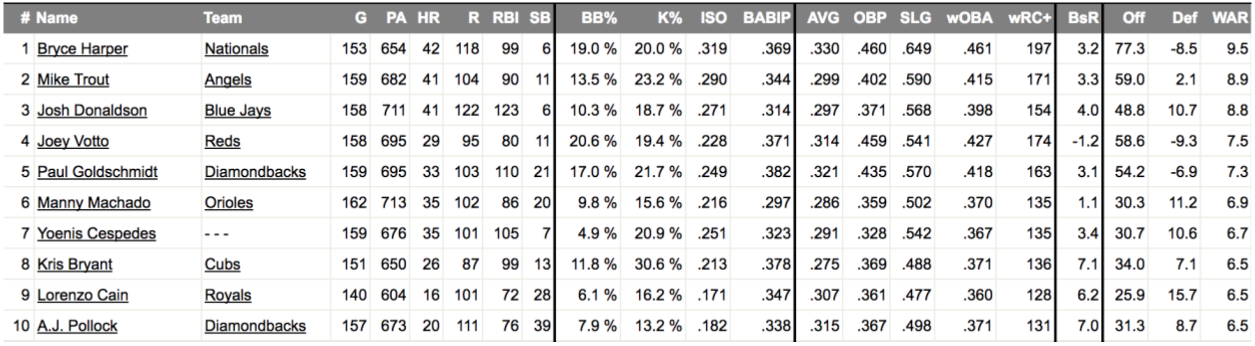
7. Most valuable offensive players according to our model

A statistic that is most frequently used in baseball analytics to summarize a player's performance in one number is WAR, or Wins Above Replacement. It attempts to measure how many more wins a player has been worth over if a replacement-level player had played instead. Obviously, WAR involves many more attributes and methodologies; however, we will attempt to use our run-producing model to pick out the top 5 run producers in 2015.



```
        player_id   h double triple hr  bb  so sb cs hbp sf top5_preds
100422 harpebr03 172     38      1 42 124 131  6  4   5  4  -225.5644
100345 goldspa01 182     38      2 33 118 151 21  5   2  7  -228.0447
100210 donaljo02 184     41      2 41  73 133  6  0   6 10  -229.1752
101211 troutmi01 172     32      6 41  92 158 11  7  10  5  -231.9484
101263 vottojo01 171     33      2 29 143 135 11  3   5  2  -234.4341
```

- The number on the very right column shows our value for run-creation in 2015. The values are negative because our original model had an intercept that was based on a team-scale. This does not affect our analysis of "who were the best run-creators".
- The run-producing leaders in 2015 according to our model were Bryce Harper, Paul Goldschmidt, Josh Donaldson, Mike Trout, and Joey Votto. Now let's compare this result to Fangraphs'(leading baseball analytics website) WAR.



#	Name	Team	G	PA	HR	R	RBI	SB	BB%	K%	ISO	BABIP	AVG	OBP	SLG	wOBA	wRC+	BsR	Off	Def	WAR
1	Bryce Harper	Nationals	153	654	42	118	99	6	19.0 %	20.0 %	.319	.369	.330	.460	.649	.461	197	3.2	77.3	-8.5	9.5
2	Mike Trout	Angels	159	682	41	104	90	11	13.5 %	23.2 %	.290	.344	.299	.402	.590	.415	171	3.3	59.0	2.1	8.9
3	Josh Donaldson	Blue Jays	158	711	41	122	123	6	10.3 %	18.7 %	.271	.314	.297	.371	.568	.398	154	4.0	48.8	10.7	8.8
4	Joey Votto	Reds	158	695	29	95	80	11	20.6 %	19.4 %	.228	.371	.314	.459	.541	.427	174	-1.2	58.6	-9.3	7.5
5	Paul Goldschmidt	Diamondbacks	159	695	33	103	110	21	17.0 %	21.7 %	.249	.382	.321	.435	.570	.418	163	3.1	54.2	-6.9	7.3
6	Manny Machado	Orioles	162	713	35	102	86	20	9.8 %	15.6 %	.216	.297	.286	.359	.502	.370	135	1.1	30.3	11.2	6.9
7	Yoenis Cespedes	- - -	159	676	35	101	105	7	4.9 %	20.9 %	.251	.323	.291	.328	.542	.367	135	3.4	30.7	10.6	6.7
8	Kris Bryant	Cubs	151	650	26	87	99	13	11.8 %	30.6 %	.213	.378	.275	.369	.488	.371	136	7.1	34.0	7.1	6.5
9	Lorenzo Cain	Royals	140	604	16	101	72	28	6.1 %	16.2 %	.171	.347	.307	.361	.477	.360	128	6.2	25.9	15.7	6.5
10	A.J. Pollock	Diamondbacks	157	673	20	111	76	39	7.9 %	13.2 %	.182	.338	.315	.367	.498	.371	131	7.0	31.3	8.7	6.5

- The leaders in WAR for 2015 were Bryce Harper, Mike Trout, Josh Donaldson, Joey Votto, and Paul Goldschmidt! Our model has successfully identified the same top 5 players!
- Obviously, our model is not perfect: for one, it does not account for defensive(fielding) ability, so it will most likely miss out on players like Lorenzo Cain, whose value is more dependent on defensive capability.
- However, if we toggle the "off"(offensive) category above and compare the ranks of players to our predictions, we are confident that we will see similar lists of players. This result demonstrates that with more advanced data and a precisely defined goal, we will be able to develop a model that better caters to our objective.

TEAM PITCHING

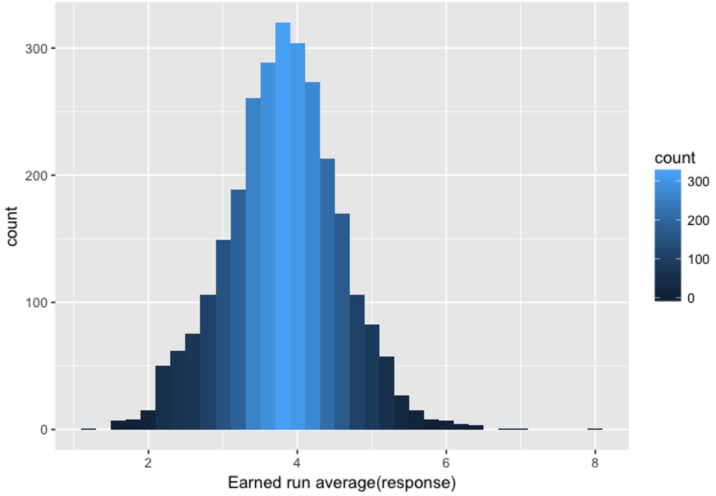
Looking at pitching stats, we are interested in examining the effect that the following stats have on a team's primary pitching stat, ERA or earned run average, which is the average number of runs allowed by a team's pitchers per nine innings.

- HA or Hits Allowed, which is the number of hits each team's pitchers allowed the opposing team's batters to get.
- HRA or Home Runs Allowed, which is the number of home runs each team's pitchers allowed the opposing team's batters to get.

- BBA or Walks allowed, which is the number of walks each team's pitchers allowed the opposing team's batters to get.
- SOA or strikeouts allowed, which is the number of strikeouts each team's pitchers got by ensuring that the opposing team's batters struck out.

1.Exploratory Data Analysis (Distribution of response data)

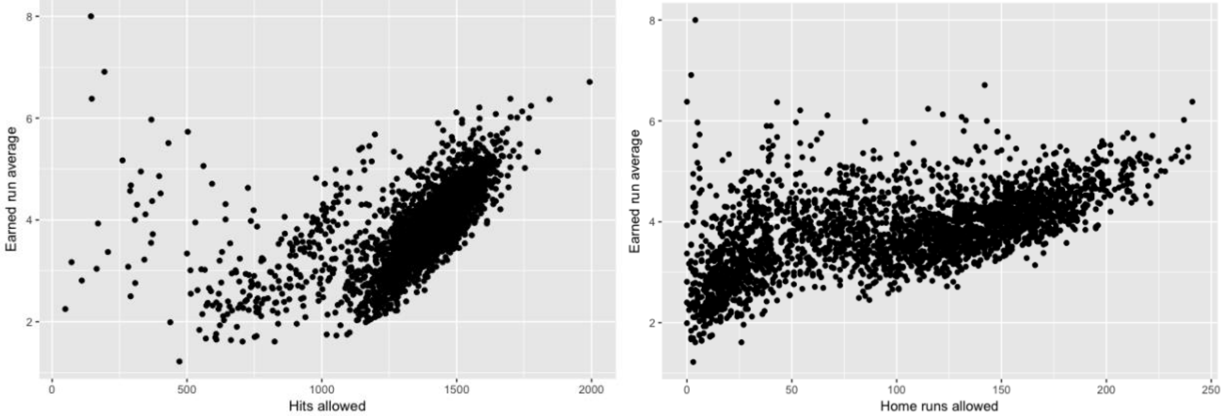
A histogram of the response variable is shown below:



The histogram tells us that the response variable is approximately normally distributed.

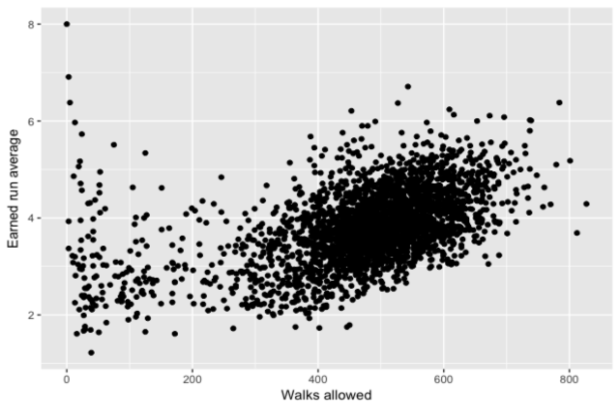
2.Exploratory Data Analysis (Scatterplots)

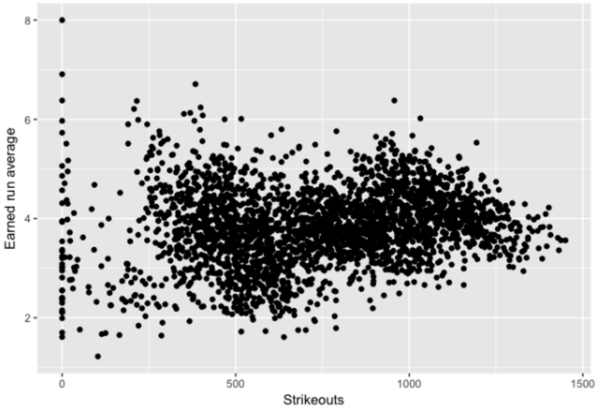
A scatter plot of the response variable (ERA) against hits allowed (HA) and home runs allowed (HRA) is shown below:



We can see that the relationship between HA and ERA is approximately linear with some non-linearity towards the left edge of the plot. We can also see that the relationship between HRA and ERA is approximately linear.

A scatter plot of the response variable (ERA) against home walks allowed (BBA) and strikeouts (SOA) is shown below:

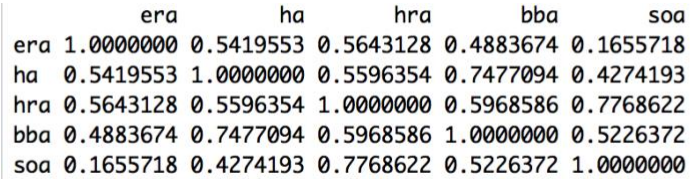




We can see that the relationship between BBA and ERA is approximately linear. We can also see that the relationship between SOA and ERA is approximately linear.

3.Exploratory Data Analysis (Correlation)

The correlation matrix between ERA and the predicting variables can be seen below:

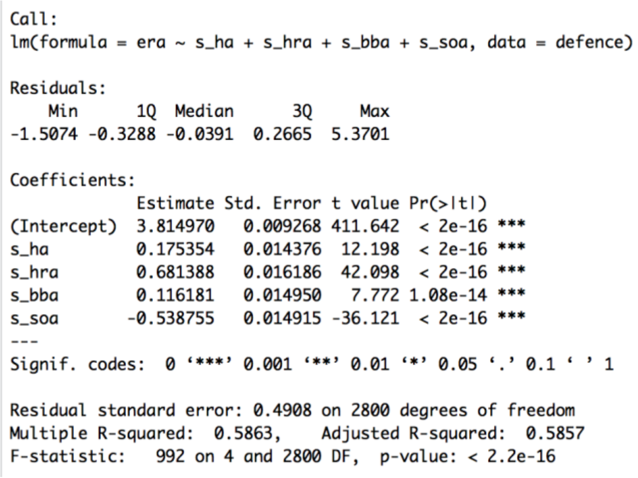


```
          era        ha       hra       bba       soa
era 1.0000000 0.5419553 0.5643128 0.4883674 0.1655718
ha  0.5419553 1.0000000 0.5596354 0.7477094 0.4274193
hra 0.5643128 0.5596354 1.0000000 0.5968586 0.7768622
bba 0.4883674 0.7477094 0.5968586 1.0000000 0.5226372
soa 0.1655718 0.4274193 0.7768622 0.5226372 1.0000000
```

Most of the predicting variables have a moderately strong correlation with the response (ERA). Some of the predictors show moderate correlation amongst one another, hence we must check the Variance Inflation Factor for multicollinearity of the model.

4.Multiple Linear Regression model

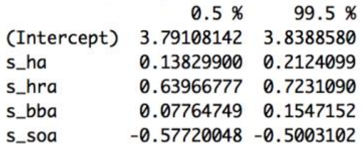
Since we see a linear relationship between the predictors and the response in the exploratory data analysis section, we shall fit a multiple linear regression model after scaling the predictors as follows:



```
Call:
lm(formula = era ~ s_ha + s_hra + s_bba + s_soa, data = defence)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.5074 -0.3288 -0.0391  0.2665  5.3701 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.814970   0.009268 411.642  < 2e-16 ***
s_ha         0.175354   0.014376  12.198  < 2e-16 ***
s_hra        0.681388   0.016186  42.098  < 2e-16 ***
s_bba        0.116181   0.014950   7.772 1.08e-14 ***
s_soa       -0.538755   0.014915 -36.121  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4908 on 2800 degrees of freedom
Multiple R-squared:  0.5863,	Adjusted R-squared:  0.5857 
F-statistic:   992 on 4 and 2800 DF,  p-value: < 2.2e-16
```

All the predictors are significant at the significance level of alpha = 0.01, and since we have only 4 predictors, all of which are significant, it doesn't make sense to opt for any variable selection methods like stepwise regression or lasso. The model has an R squared value of 0.58 indicating that it explains approximately 58% of the variability in the data which is considered good.

The confidence interval for the coefficients can be seen below:

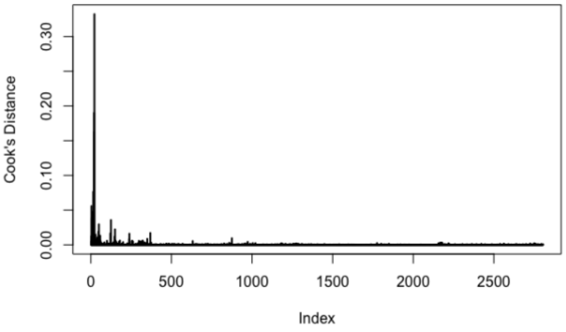


```
                  0.5 %     99.5 %
(Intercept)  3.79108142  3.8388580
s_ha         0.13829900  0.2124099
s_hra        0.63966777  0.7231090
s_bba        0.07764749  0.1547152
s_soa       -0.57720048 -0.5003102
```

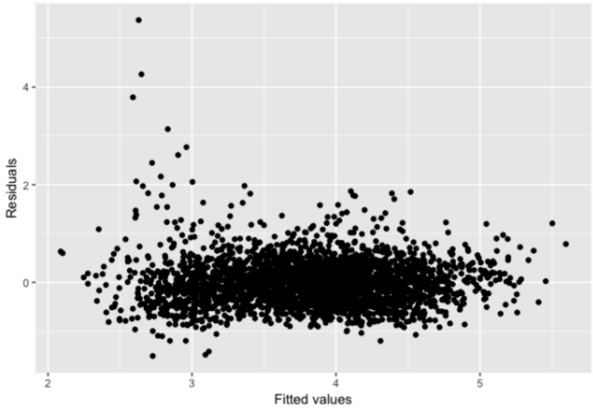
Since all the predictors are statistically significant at alpha = 0.01, the 99% confidence interval doesn't contain 0 for any of the predictors.

5.Goodness of fit analysis

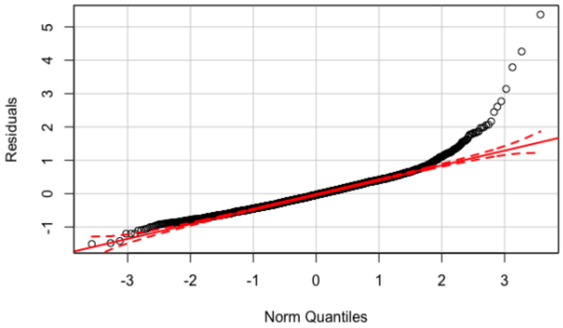
First, taking a look at the Cook's distance plot for the model, it seems that none of the points exceed a cook's distance of 0.35. Thus, we conclude that there are no outliers in the data, and there is no need to remove data points and re-run the model.

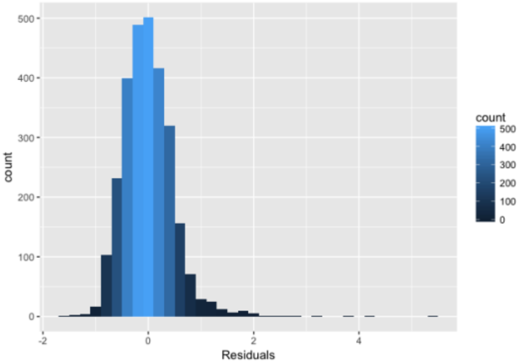


Second, taking a look at the fitted values versus residuals plot, we conclude that there is no sign of non-constant variance. Therefore, our assumption of constant variance holds.



Third, taking a look at the QQ plot and histogram of the residuals, it seems that the assumption of normality is violated at the upper quantiles, but not badly enough for us to completely discard the analysis and conclusions we draw from the model. Only a few of the 2,805 data points deviate from the expected theoretical normal quantiles. Therefore, we conclude that our assumption of normality holds approximately.





Lastly, we examine the Variance Inflation Factor (VIF) of the model to make sure that there is no multicollinearity in our model.

scaled_ha: 2.405356 ,scaled _hra: 3.049136 ,scaled _bba: 2.601122, scaled _soa: 2.589160

Since none of the VIFs for the scaled coefficients exceeds 10, we conclude that there is no multicollinearity.

6.Conclusions regarding the model for team pitching

The coefficient of the multiple linear regression model can be seen below:
scaled_ha: 0.175354
scaled_hra: 0.681388
scaled_bba: 0.116181

scaled_soa: -0.538755

We see that for every unit increase in scaled hits allowed, the earned run average increases by 0.17 units, for every unit increase in scaled home runs allowed, the earned run average increases by 0.68 units, for every unit increase in scaled walk allowed, the earned run average increase by 0.11 units and for every unit increase in strikeouts, the earned run average decreases by 0.53 units.

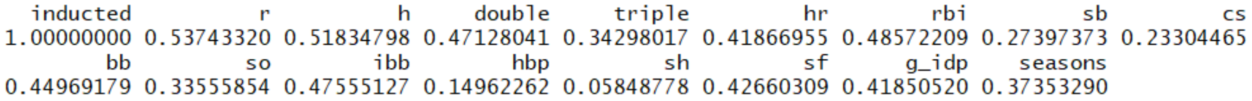
Since the goal of the pitching (defending) team is to minimize the earned run average, to maximize their chances of winning the match, it seems that the primary focus of the pitching team should be preventing the opposition from hitting a home run as it contributes to the greatest decrease in the earned run average. The secondary focus of the pitching team should be to get as many strikeouts as possible as this contributes to the second greatest decrease in the earned run average. Lastly, the pitching team should try their best to minimize the number of hits scored and walks taken by the opposing team, as this also contributes to the reduction of the earned run average.

HALL OF FAME

Looking at the hall of fame ballots and inducted players, we want to know what is the probability that a batter being in the ballot, given his batting statistics, being inducted to the baseball hall of fame from 1980 onwards. To achieve that, we used the hall of fame data that has the players that were in the ballots, the years and if they were inducted or not, and the batting dataset that contains all batting statistics of each player per year.

The dataset contained the following attributes:
inducted(binary), runs, hits, double runs, triple runs, home runs, runs batted in, stolen bases, caught stolen, walks, strike out, intentional walk, times hit by pitches, sacrifice hits, sacrifice flies, grounded into double play and seasons

1.Exploratory data analysis (correlation)



```
  inducted          r          h     double     triple         hr        rbi         sb         cs 
1.00000000 0.53743320 0.51834798 0.47128041 0.34298017 0.41866955 0.48572209 0.27397373 0.23304465 
        bb         so        ibb        hbp         sh         sf      g_idp    seasons 
0.44969179 0.33555854 0.47555127 0.14962262 0.05848778 0.42660309 0.41850520 0.37353290 
```

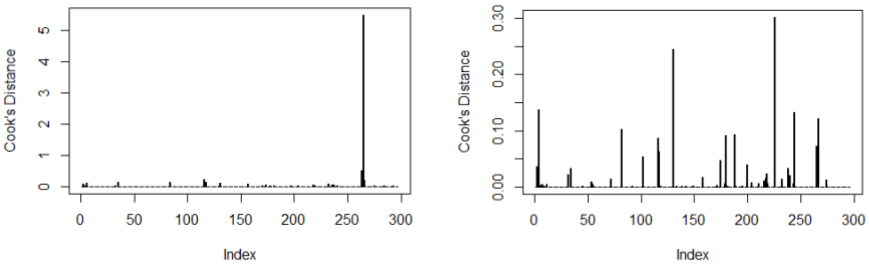
The predicting variables show low to moderate correlation with the response variable “inducted”. But some predictors show high correlation between each other. We will use variable selection to try to mitigate the correlation and vif to evaluate the multicollinearity of the final model.

2.Logistic Regression Model

Since the data has a binary response that are independent random variables, and we want to be able to predict if the response of a certain set of attributes would be 1, we decided to use a logistic regression model.

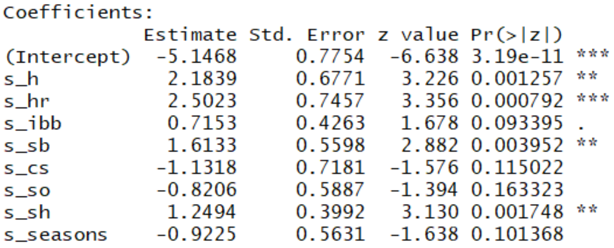
The data was standardized by columns and a training set was selected randomly, containing 80% of the observations. Used the glm function, family “binomial”, to estimate the model. Using all the initial attributes, most of the estimates were not statistically significant, even at the significance level of alpha = 0.1.

Used cook’s distance to evaluate possible outliers in the dataset. It was possible to notice an outlier that had a much greater distance than the other observations (left). After investigating the observation, decided to remove it and recalculated the cook’s distance (right).

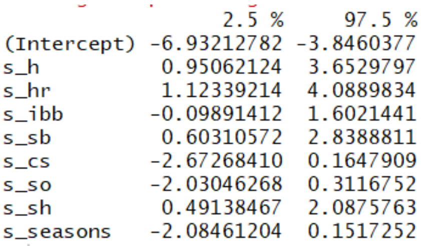


With the outlier removed, decided to use lasso and AIC to select the attributes and improve the model.

AIC showed a better performance overall compared to lasso. AIC removed eight predictors from the initial model. Even though not all of them were statistically significant, the variable selection greatly improved the results.



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -5.1468     0.7754  -6.638 3.19e-11 ***
s_h           2.1839     0.6771   3.226 0.001257 ** 
s_hr          2.5023     0.7457   3.356 0.000792 ***
s_ibb         0.7153     0.4263   1.678 0.093395 .  
s_sb          1.6133     0.5598   2.882 0.003952 ** 
s_cs         -1.1318     0.7181  -1.576 0.115022    
s_so         -0.8206     0.5887  -1.394 0.163323    
s_sh          1.2494     0.3992   3.130 0.001748 ** 
s_seasons    -0.9225     0.5631  -1.638 0.101368    
```



```
                  2.5 %     97.5 %
(Intercept) -6.93212782 -3.8460377
s_h          0.95062124  3.6529797
s_hr         1.12339214  4.0889834
s_ibb       -0.09891412  1.6021441
s_sb         0.60310572  2.8388811
s_cs        -2.67268410  0.1647909
s_so        -2.03046268  0.3116752
s_sh         0.49138467  2.0875763
s_seasons   -2.08461204  0.1517252
```

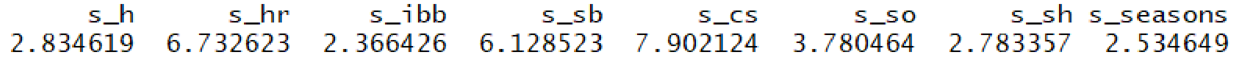
3.Goodness of fit analysis

Since our dataset didn’t have replications, decided to use the Hosmer-Lemeshow test to assess if the model was a good fit. Ran the test for groups from 5 to 15, with the following p-value results:

[1] 0.8090245 [1] 0.9585331 [1] 0.9605069 [1] 0.9543258 [1] 0.9418811 [1] 0.9631239
[1] 0.9803991 [1] 0.9934196 [1] 0.9971449 [1] 0.9977394 [1] 0.8368607

The values were large, not sufficient to reject the null hypothesis, and we then concluded that the model was a good fit.

We then did a multicollinearity analysis, since the predictors initially showed signs of high correlation.

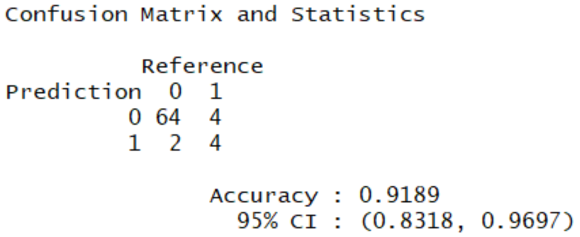


```
      s_h       s_hr      s_ibb       s_sb       s_cs       s_so       s_sh s_seasons 
 2.834619  6.732623  2.366426  6.128523  7.902124  3.780464  2.783357  2.534649 
```

The results for all the predictors were below 10, so the impact of multicollinearity is not strong for the final model. Variable selection definitely improved the properties of the model.

4.Conclusions

To evaluate the accuracy of the model, probabilities greater than 0.5 were classified as 1 and lower or equal to 0.5 were classified as 0. The results of the model in the test set were:



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 64  4
         1  2  4

               Accuracy : 0.9189
                 95% CI : (0.8318, 0.9697)
```

The model classified 91.89% of the test set correctly. The greatest source of error was from false negatives.

We initially thought that the number of runs were the most important factor for a player to be inducted to the hall of fame. AIC removed runs from the model, and we believe that it happened because hits and runs are highly correlated. Home runs has the highest impact on the log odds, given an increase of scaled unit holding all the other predictors constant.

Categories that indicate flaws in players games, caught stolen bases and strike out, have a negative influence on the probability of being inducted to the hall of fame, and that confirms our initial analysis.

We concluded that the most important statistics for a batter to be inducted to the hall of fame, according to the historical data, are home runs, hits, stolen bases and sacrifice hits.

FURTHER INQUIRIES

For this project, we have explored Sean Lahman's database and the generic baseball statistics that were provided to build our models. However, current advances in baseball analytics reach far beyond such basic stats; there are libraries of different, specified, and context-specific databases on the web that are accessible for further analysis. We hope to incorporate such statistics into our analysis and build a more precise model in the future.