Data 621 Business Analytics and Data Mining
HW1
Xiaomeng(Vivian) Kong

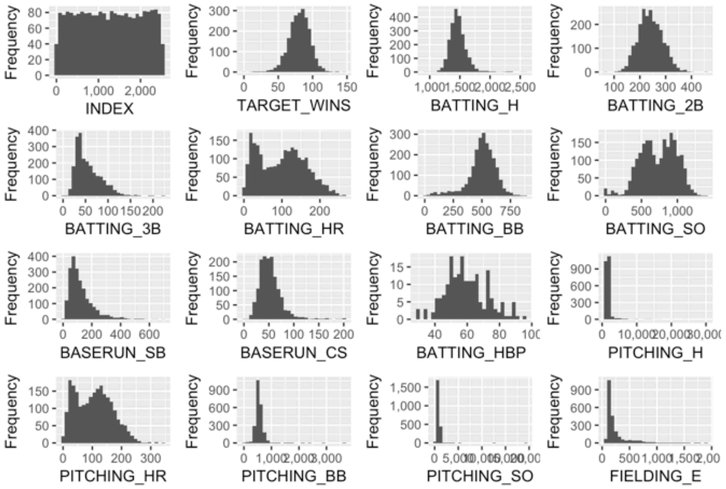
6/11/2018

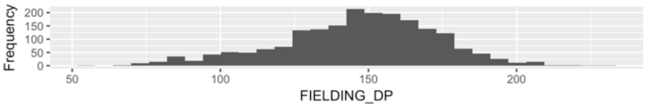
1. Data Exploration

In our training data, I find 2276 rows and 16 columns. Each rows represent a professional baseball team's performance from the year 1871 to 2006 inclusive. We're interested in finding the trend of number of wins and predict them by using multiple linear regression model on the training data for the team. To find the model, we need to explore each variable first to get the better idea about our data.

- The shape of the distribution.

I used the histogram for each variables. From the histogram (Except the column of "Index"), I made a chart to categorize the shape of the distributions of all 15 variables



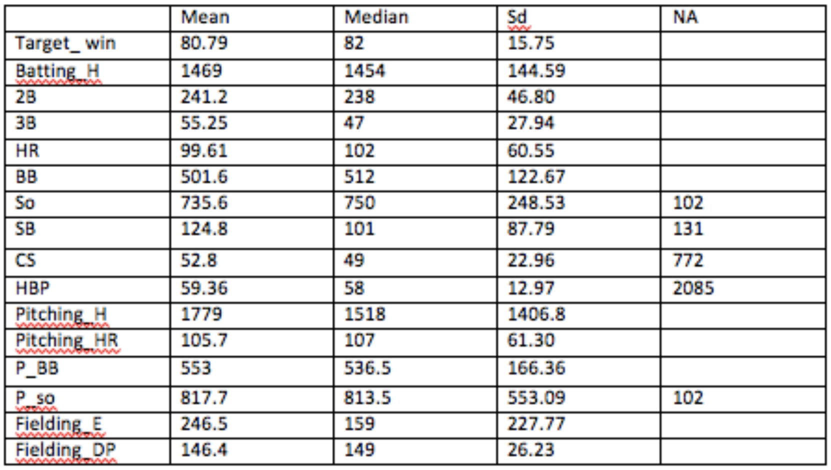


Shape of the distribution	Variables
Approximately normal	Target_Wins, Batting_H,Pitching_BB
Skewed Right	Batting_3B, Baserun_3B, Baserun_SB,Baserun_CS, Pitching_H, Fielding_E
Skewed left	Batting_BB, Fielding_DP

Bi-modal	Battting_2B, Batting_HR, Batting_SO, Pitching_HR
Abnormal shape	Batting_HBP, Pitching_SO

For the abnormal shape of distributions, we will suspect that something wrong with those variables. Before we move on to give the conclusion, I found the central tendency and standard deviation for these 17 variables.

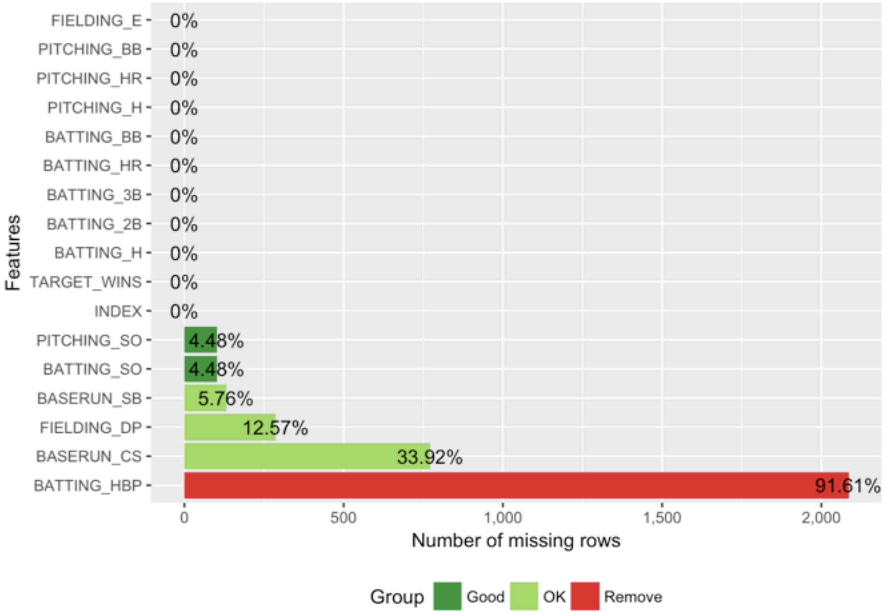
- Central Tendency and Standard Deviation



	Mean	Median	Sd	NA
Target_ win	80.79	82	15.75	
Batting_H	1469	1454	144.59	
2B	241.2	238	46.80	
3B	55.25	47	27.94	
HR	99.61	102	60.55	
BB	501.6	512	122.67	
So	735.6	750	248.53	102
SB	124.8	101	87.79	131
CS	52.8	49	22.96	772
HBP	59.36	58	12.97	2085
Pitching_H	1779	1518	1406.8	
Pitching_HR	105.7	107	61.30	
P_BB	553	536.5	166.36	
P_so	817.7	813.5	553.09	102
Fielding_E	246.5	159	227.77	
Fielding_DP	146.4	149	26.23	

- Missing values for the variables

There are 5 variables that have missing data set.



I replace the median for the missing values for these 4 variables except Batting_HBP. Batting_HBP missed 91.61% of the data values which should be removed. The reason why I chose the median for the other 4 variables because median is a better indicator for the central tendency, compare to the mean

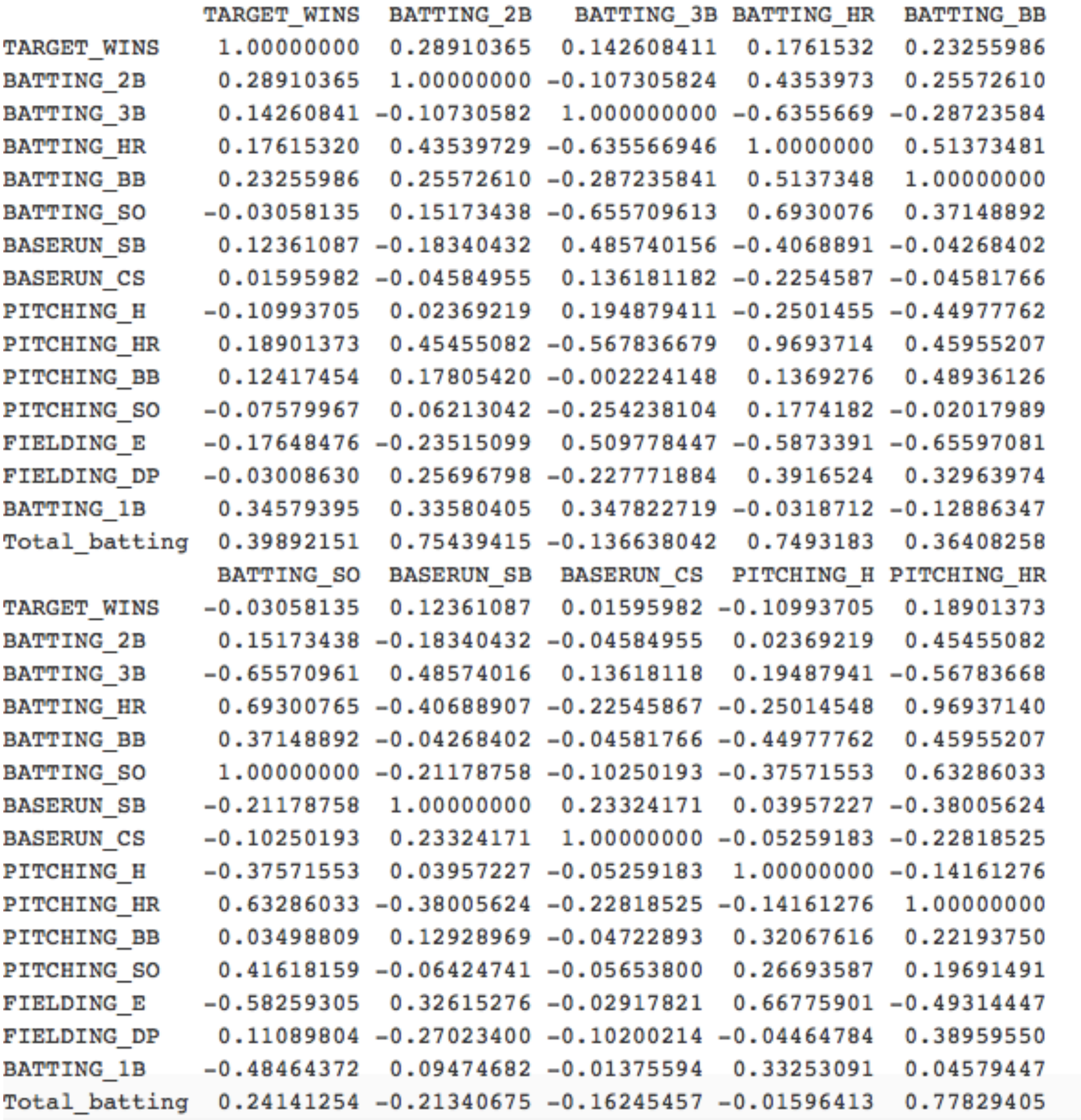
Because the median is not affected by extreme values.

- New Variables

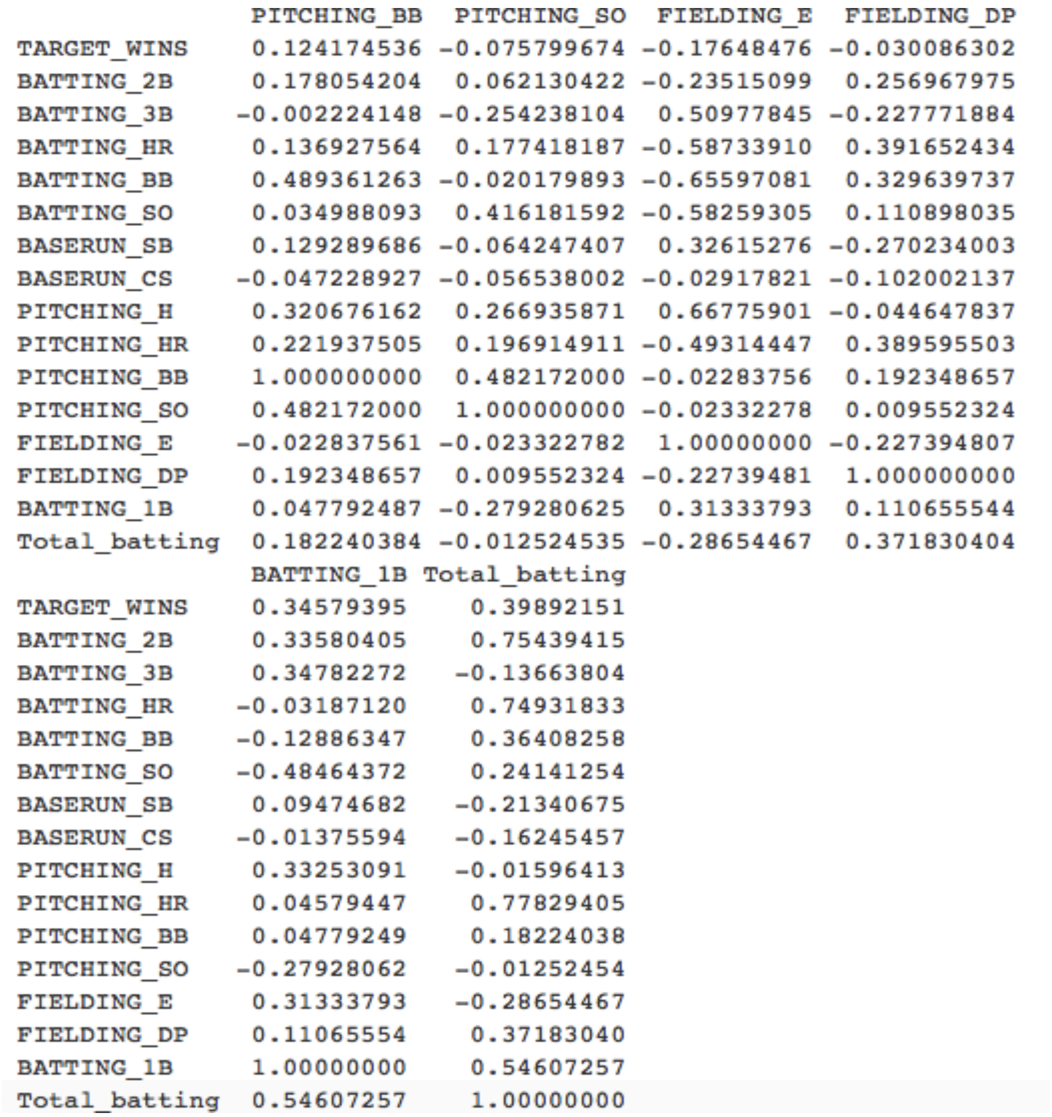
I find out that the Batting_H=Batting_1B+ Batting_2B+ Batting_3B+Batting_HR, so in this case, I create the new variable Batting_1B= Batting_H- Batting_2B-Batting_3B-Batting_HR.
Another new variable is Total_Batting= 1*Batting_1B+2*Batting_2B+3*Batting_3B+4*Batting_HR.

- Correlations

After I fixed the missing values, I can find the correlations between variables. I also add the new variable BATTING_1B and Total_batting in the chart.



```
              TARGET_WINS  BATTING_2B   BATTING_3B BATTING_HR   BATTING_BB
TARGET_WINS    1.00000000  0.28910365  0.142608411  0.1761532  0.23255986
BATTING_2B     0.28910365  1.00000000 -0.107305824  0.4353973  0.25572610
BATTING_3B     0.14260841 -0.10730582  1.000000000 -0.6355669 -0.28723584
BATTING_HR     0.17615320  0.43539729 -0.635566946  1.0000000  0.51373481
BATTING_BB     0.23255986  0.25572610 -0.287235841  0.5137348  1.00000000
BATTING_SO    -0.03058135  0.15173438 -0.655709613  0.6930076  0.37148892
BASERUN_SB     0.12361087 -0.18340432  0.485740156 -0.4068891 -0.04268402
BASERUN_CS     0.01595982 -0.04584955  0.136181182 -0.2254587 -0.04581766
PITCHING_H    -0.10993705  0.02369219  0.194879411 -0.2501455 -0.44977762
PITCHING_HR    0.18901373  0.45455082 -0.567836679  0.9693714  0.45955207
PITCHING_BB    0.12417454  0.17805420 -0.002224148  0.1369276  0.48936126
PITCHING_SO   -0.07579967  0.06213042 -0.254238104  0.1774182 -0.02017989
FIELDING_E    -0.17648476 -0.23515099  0.509778447 -0.5873391 -0.65597081
FIELDING_DP   -0.03008630  0.25696798 -0.227771884  0.3916524  0.32963974
BATTING_1B     0.34579395  0.33580405  0.347822719 -0.0318712 -0.12886347
Total_batting  0.39892151  0.75439415 -0.136638042  0.7493183  0.36408258
               BATTING_SO  BASERUN_SB  BASERUN_CS  PITCHING_H PITCHING_HR
TARGET_WINS   -0.03058135  0.12361087  0.01595982 -0.10993705  0.18901373
BATTING_2B     0.15173438 -0.18340432 -0.04584955  0.02369219  0.45455082
BATTING_3B    -0.65570961  0.48574016  0.13618118  0.19487941 -0.56783668
BATTING_HR     0.69300765 -0.40688907 -0.22545867 -0.25014548  0.96937140
BATTING_BB     0.37148892 -0.04268402 -0.04581766 -0.44977762  0.45955207
BATTING_SO     1.00000000 -0.21178758 -0.10250193 -0.37571553  0.63286033
BASERUN_SB    -0.21178758  1.00000000  0.23324171  0.03957227 -0.38005624
BASERUN_CS    -0.10250193  0.23324171  1.00000000 -0.05259183 -0.22818525
PITCHING_H    -0.37571553  0.03957227 -0.05259183  1.00000000 -0.14161276
PITCHING_HR    0.63286033 -0.38005624 -0.22818525 -0.14161276  1.00000000
PITCHING_BB    0.03498809  0.12928969 -0.04722893  0.32067616  0.22193750
PITCHING_SO    0.41618159 -0.06424741 -0.05653800  0.26693587  0.19691491
FIELDING_E    -0.58259305  0.32615276 -0.02917821  0.66775901 -0.49314447
FIELDING_DP    0.11089804 -0.27023400 -0.10200214 -0.04464784  0.38959550
BATTING_1B    -0.48464372  0.09474682 -0.01375594  0.33253091  0.04579447
Total_batting  0.24141254 -0.21340675 -0.16245457 -0.01596413  0.77829405
```



```
              PITCHING_BB  PITCHING_SO  FIELDING_E  FIELDING_DP
TARGET_WINS    0.124174536 -0.075799674 -0.17648476 -0.030086302
BATTING_2B     0.178054204  0.062130422 -0.23515099  0.256967975
BATTING_3B    -0.002224148 -0.254238104  0.50977845 -0.227771884
BATTING_HR     0.136927564  0.177418187 -0.58733910  0.391652434
BATTING_BB     0.489361263 -0.020179893 -0.65597081  0.329639737
BATTING_SO     0.034988093  0.416181592 -0.58259305  0.110898035
BASERUN_SB     0.129289686 -0.064247407  0.32615276 -0.270234003
BASERUN_CS    -0.047228927 -0.056538002 -0.02917821 -0.102002137
PITCHING_H     0.320676162  0.266935871  0.66775901 -0.044647837
PITCHING_HR    0.221937505  0.196914911 -0.49314447  0.389595503
PITCHING_BB    1.000000000  0.482172000 -0.02283756  0.192348657
PITCHING_SO    0.482172000  1.000000000 -0.02332278  0.009552324
FIELDING_E    -0.022837561 -0.023322782  1.00000000 -0.227394807
FIELDING_DP    0.192348657  0.009552324 -0.22739481  1.000000000
BATTING_1B     0.047792487 -0.279280625  0.31333793  0.110655544
Total_batting  0.182240384 -0.012524535 -0.28654467  0.371830404
              BATTING_1B Total_batting
TARGET_WINS    0.34579395    0.39892151
BATTING_2B     0.33580405    0.75439415
BATTING_3B     0.34782272   -0.13663804
BATTING_HR    -0.03187120    0.74931833
BATTING_BB    -0.12886347    0.36408258
BATTING_SO    -0.48464372    0.24141254
BASERUN_SB     0.09474682   -0.21340675
BASERUN_CS    -0.01375594   -0.16245457
PITCHING_H     0.33253091   -0.01596413
PITCHING_HR    0.04579447    0.77829405
PITCHING_BB    0.04779249    0.18224038
PITCHING_SO   -0.27928062   -0.01252454
FIELDING_E     0.31333793   -0.28654467
FIELDING_DP    0.11065554    0.37183040
BATTING_1B     1.00000000    0.54607257
Total_batting  0.54607257    1.00000000
```

Between Target_wins and other variables, there are correlations but not strong. Most of them are showing weak positive/negative correlations. It means that when we set up the multi regression model later, we will not expect high R^2.

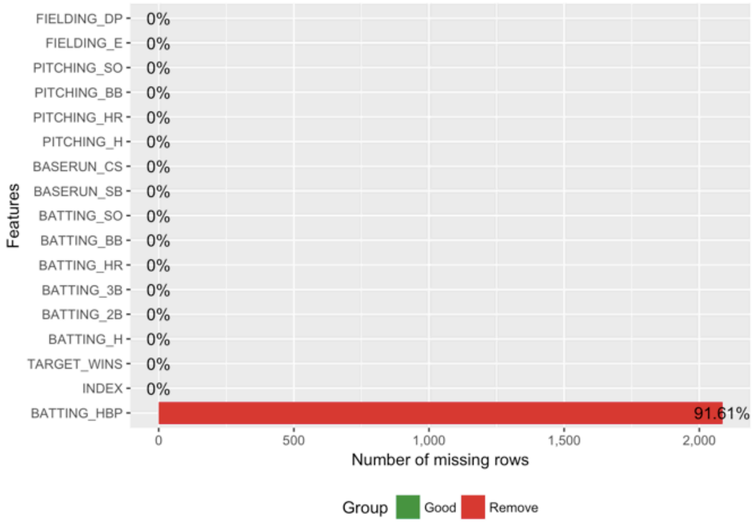
Between variables, I found that Batting_HR highly correlated to Pitching_HR. (Unfortunately, I don't really understand the Baseball. I just guess the following correlation. Because there is the same group of players batting and pitching in turns for the same game. I think that is the reason why they are very correlated.

2. Data Preparation

- Fixing the missing values

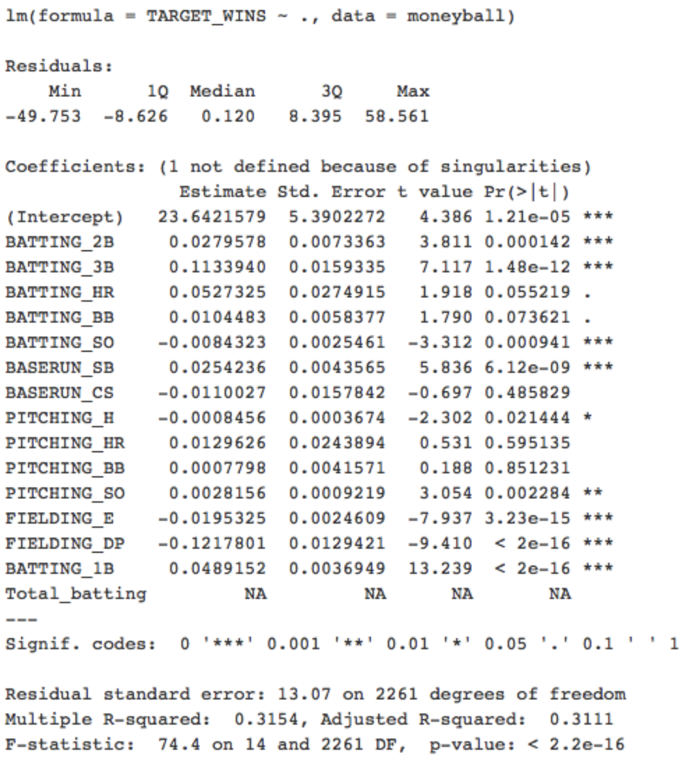
  As I did from part 1. Data exploration, I removed the "Batting_HBP" Since it lacks about 92% of data values so it is not necessary to fix it.

For Pitching_So, Batting_So, Baserun_SB, Baserun_CS, Fielding_DP, I used the median to replace the missing values since the median is a better indicator for the central measurement. The median is not affected by the extreme value.

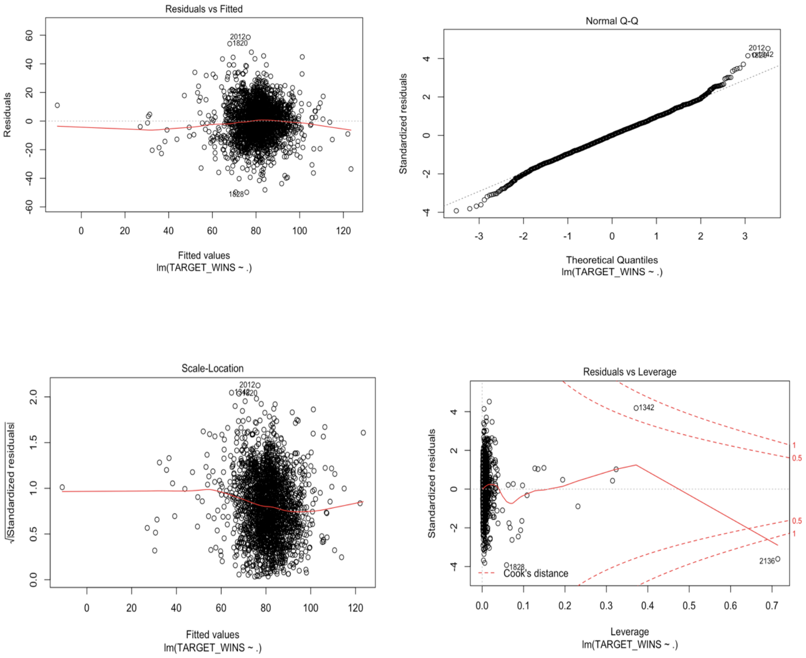


- Combing variables.
  I combined the Batting_H, Batting_2B, Batting_3B, Batting_HR and create the new variable: Batting_1B.
  I also created another new variable:
  Total_Batting=1*Batting_1B+2*Batting_2B+3*Batting_3B+4*Batting_HR

3. Building Models
   1) Model 1
      Backward Selections. For the first model, I will use all variables in the models, and then decide which one variable I will eliminate.



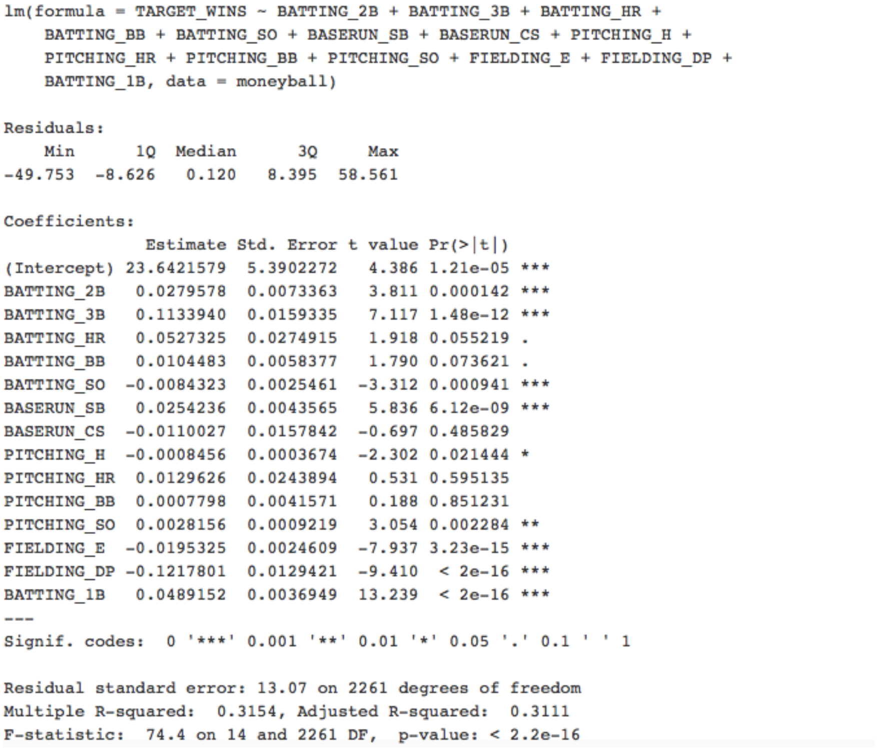
```
lm(formula = TARGET_WINS ~ ., data = moneyball)

Residuals:
    Min      1Q  Median      3Q     Max 
-49.753  -8.626   0.120   8.395  58.561 

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   23.6421579  5.3902272   4.386 1.21e-05 ***
BATTING_2B     0.0279578  0.0073363   3.811 0.000142 ***
BATTING_3B     0.1133940  0.0159335   7.117 1.48e-12 ***
BATTING_HR     0.0527325  0.0274915   1.918 0.055219 .  
BATTING_BB     0.0104483  0.0058377   1.790 0.073621 .  
BATTING_SO    -0.0084323  0.0025461  -3.312 0.000941 ***
BASERUN_SB     0.0254236  0.0043565   5.836 6.12e-09 ***
BASERUN_CS    -0.0110027  0.0157842  -0.697 0.485829    
PITCHING_H    -0.0008456  0.0003674  -2.302 0.021444 *  
PITCHING_HR    0.0129626  0.0243894   0.531 0.595135    
PITCHING_BB    0.0007798  0.0041571   0.188 0.851231    
PITCHING_SO    0.0028156  0.0009219   3.054 0.002284 ** 
FIELDING_E    -0.0195325  0.0024609  -7.937 3.23e-15 ***
FIELDING_DP   -0.1217801  0.0129421  -9.410  < 2e-16 ***
BATTING_1B     0.0489152  0.0036949  13.239  < 2e-16 ***
Total_batting         NA         NA      NA       NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.07 on 2261 degrees of freedom
Multiple R-squared:  0.3154, Adjusted R-squared:  0.3111 
F-statistic:  74.4 on 14 and 2261 DF,  p-value: < 2.2e-16
```



From this model, we get R^2 is 0.3154 which means that 31.54% of the total variance of Target_wins can be explained by the total variance of all variables. It is not quite a satisfying number. F statistics tell us if the model best fits the population from which the data are sampled. For our case, we get p-value is approximately 0 and it is statistically significant. It indicates that our model fits the population from the sample we took. However, in this case, we have "NA" in the Total_batting which means it may strongly correlated to some other independent variables.

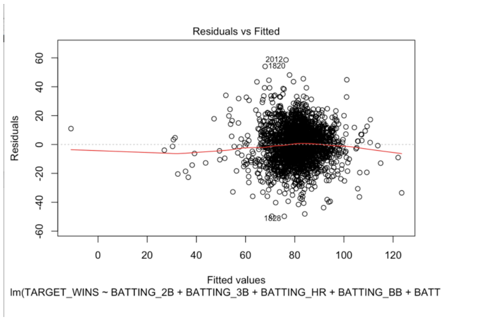
2) Model 2

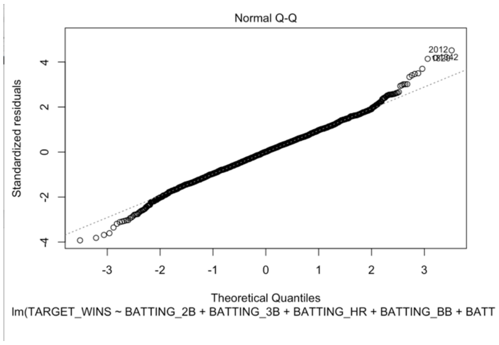
I got rid of “Total_batting” and take a look at the linear model again.

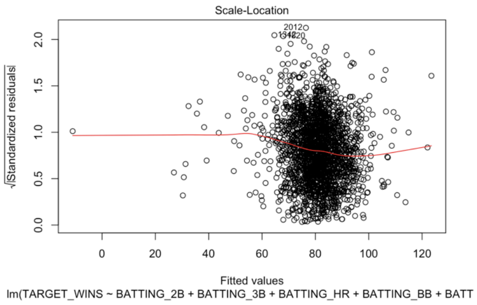


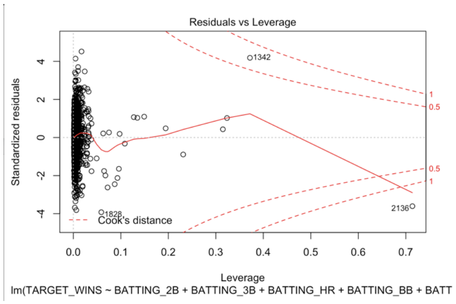
```
lm(formula = TARGET_WINS ~ BATTING_2B + BATTING_3B + BATTING_HR +
    BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
    PITCHING_HR + PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP +
    BATTING_1B, data = moneyball)

Residuals:
    Min      1Q  Median      3Q     Max 
-49.753  -8.626   0.120   8.395  58.561 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23.6421579  5.3902272   4.386 1.21e-05 ***
BATTING_2B   0.0279578  0.0073363   3.811 0.000142 ***
BATTING_3B   0.1133940  0.0159335   7.117 1.48e-12 ***
BATTING_HR   0.0527325  0.0274915   1.918 0.055219 .  
BATTING_BB   0.0104483  0.0058377   1.790 0.073621 .  
BATTING_SO  -0.0084323  0.0025461  -3.312 0.000941 ***
BASERUN_SB   0.0254236  0.0043565   5.836 6.12e-09 ***
BASERUN_CS  -0.0110027  0.0157842  -0.697 0.485829    
PITCHING_H  -0.0008456  0.0003674  -2.302 0.021444 *  
PITCHING_HR  0.0129626  0.0243894   0.531 0.595135    
PITCHING_BB  0.0007798  0.0041571   0.188 0.851231    
PITCHING_SO  0.0028156  0.0009219   3.054 0.002284 ** 
FIELDING_E  -0.0195325  0.0024609  -7.937 3.23e-15 ***
FIELDING_DP -0.1217801  0.0129421  -9.410  < 2e-16 ***
BATTING_1B   0.0489152  0.0036949  13.239  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.07 on 2261 degrees of freedom
Multiple R-squared:  0.3154,	Adjusted R-squared:  0.3111 
F-statistic:  74.4 on 14 and 2261 DF,  p-value: < 2.2e-16
```

In this case, we get the R^2 which is the same as the first case.



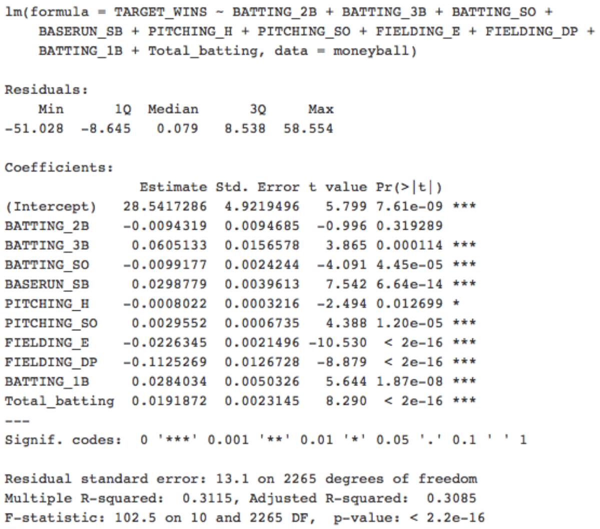




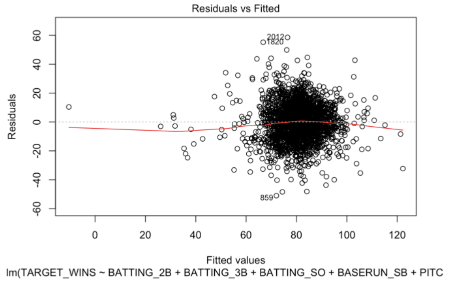


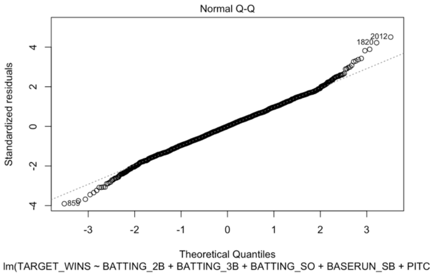
3) Model 3

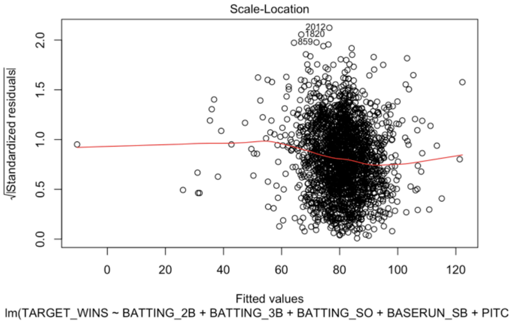
I got rid of the variables that are not statistically significant. In this model, I got rid of “Batting_HR, Batting_BB, Baserun_CS, Pitching_HR, and Pitching_BB.

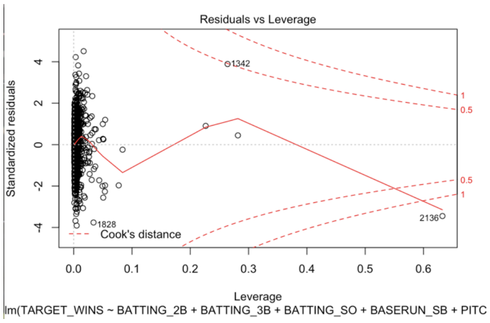


```
lm(formula = TARGET_WINS ~ BATTING_2B + BATTING_3B + BATTING_SO +
    BASERUN_SB + PITCHING_H + PITCHING_SO + FIELDING_E + FIELDING_DP +
    BATTING_1B + Total_batting, data = moneyball)

Residuals:
    Min      1Q  Median      3Q     Max
-51.028  -8.645   0.079   8.538  58.554

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   28.5417286  4.9219496   5.799 7.61e-09 ***
BATTING_2B    -0.0094319  0.0094685  -0.996 0.319289
BATTING_3B     0.0605133  0.0156578   3.865 0.000114 ***
BATTING_SO    -0.0099177  0.0024244  -4.091 4.45e-05 ***
BASERUN_SB     0.0298779  0.0039613   7.542 6.64e-14 ***
PITCHING_H    -0.0008022  0.0003216  -2.494 0.012699 *
PITCHING_SO    0.0029552  0.0006735   4.388 1.20e-05 ***
FIELDING_E    -0.0226345  0.0021496 -10.530  < 2e-16 ***
FIELDING_DP   -0.1125269  0.0126728  -8.879  < 2e-16 ***
BATTING_1B     0.0284034  0.0050326   5.644 1.87e-08 ***
Total_batting  0.0191872  0.0023145   8.290  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.1 on 2265 degrees of freedom
Multiple R-squared:  0.3115, Adjusted R-squared:  0.3085
F-statistic: 102.5 on 10 and 2265 DF,  p-value: < 2.2e-16
```



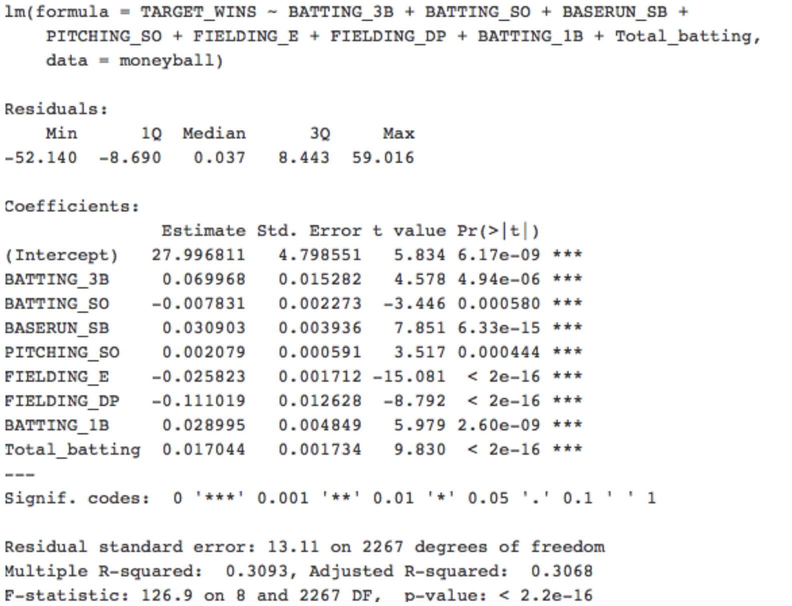




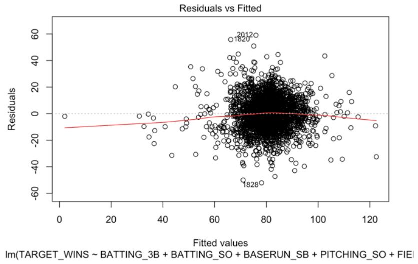


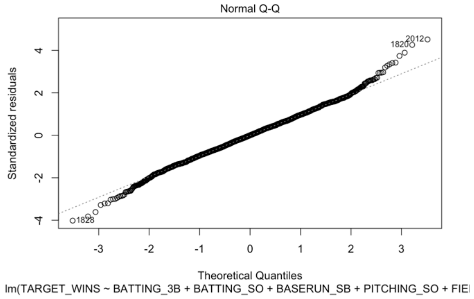
4) Model 4

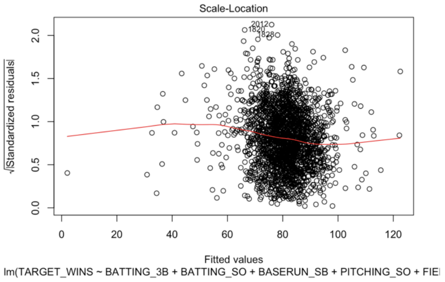
Keep getting rid of the relatively statistically insignificant variable. Batting_2B and Piting_H is being got rid of.

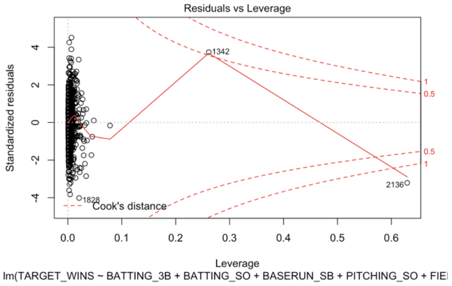


```
lm(formula = TARGET_WINS ~ BATTING_3B + BATTING_SO + BASERUN_SB +
    PITCHING_SO + FIELDING_E + FIELDING_DP + BATTING_1B + Total_batting,
    data = moneyball)

Residuals:
    Min      1Q  Median      3Q     Max
-52.140  -8.690   0.037   8.443  59.016

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   27.996811   4.798551   5.834 6.17e-09 ***
BATTING_3B     0.069968   0.015282   4.578 4.94e-06 ***
BATTING_SO    -0.007831   0.002273  -3.446 0.000580 ***
BASERUN_SB     0.030903   0.003936   7.851 6.33e-15 ***
PITCHING_SO    0.002079   0.000591   3.517 0.000444 ***
FIELDING_E    -0.025823   0.001712 -15.081  < 2e-16 ***
FIELDING_DP   -0.111019   0.012628  -8.792  < 2e-16 ***
BATTING_1B     0.028995   0.004849   5.979 2.60e-09 ***
Total_batting  0.017044   0.001734   9.830  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.11 on 2267 degrees of freedom
Multiple R-squared:  0.3093, Adjusted R-squared:  0.3068
F-statistic: 126.9 on 8 and 2267 DF,  p-value: < 2.2e-16
```









4. Select Models

Assume that the model with the highest adjusted R^2 is the “best” model especially for this case. Since all the independent variables are not strongly correlated to the Target_wins. Our Adjusted R^2 are 31.11%, 31.11%, 30.85%, 30.68%. Based on our assumption. The first model with all independent variables and the second model without “Total_batting” are the best models.

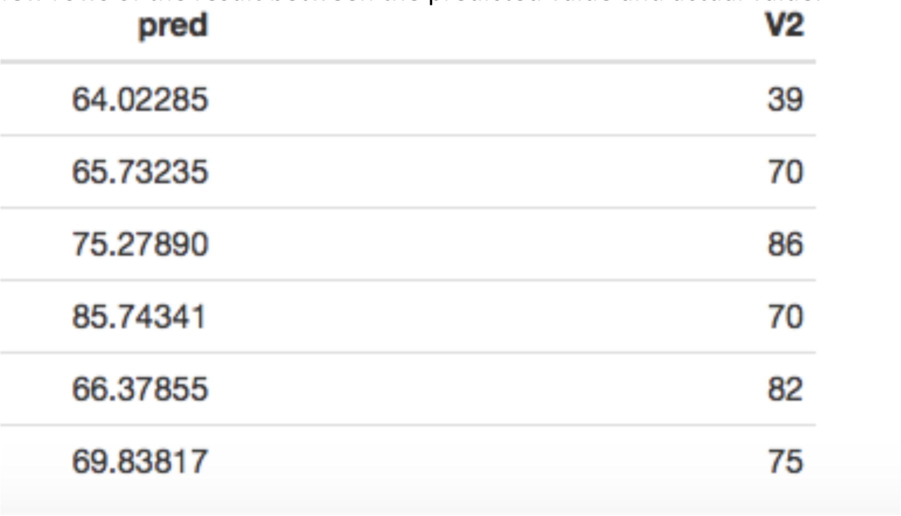
All the models, F_test gives us the result that all independent variables have correlation with the Target_wins in a strong statistically significant level since the P-value is approximately 0.

In the meanwhile, for all the models, the residual plots seem to be reasonable but not the satisfying results. The plots show that our models can predict Target_wins but the models do show the significant amount of residuals when we compare the expected value and real values. It means that the model still need a lot of improvement. It requires more variables adjustment. (Unfortunately, I have no idea about the baseball, so a lot of technical terms I felt confused with, even with a lot of “Wikipedia” learning.) It also requires that more data collections because we have a lot of missing values and values probably outdated as well.

For selecting models, I am going to use the highest adjusted R^2 although I felt the worse model that makes sense. The “Worse” model may be not that “worse”, it just needs more data collections and more manipulation for the variables.

Prediction

I chose the first model to make the prediction. In order to make the prediction, I fixed my “Evaluation” data set to fit my model, I added Batting_Base 1 and Total_batting, removed the Batting_HBP, and replaced the median to the missing values. The following pic is just the first a few rows of the result between the predicted value and actual value.



pred	V2
64.02285	39
65.73235	70
75.27890	86
85.74341	70
66.37855	82
69.83817	75

Appendix

Github: https://github.com/xkong100/data-621/blob/master/Hw1/XKong_Assignment1_data621.Rmd