DATA 621 – Business Analytics and Data Mining

Homework #1 Assignment Requirements

Alain Kuiete Tchoupou

INTRODUCTION

Study of 2276 professionals baseball teams from 1871 to 2006. There are 16 columns where 15 are predictors.

DATA EXPLORATION

This table gives the definition of each variables

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	Positive Impact on Wins
TEAM_BATTING_H	Base Hits by batters(1B,2B,3B,HR)	Positive Impact on Wins
TEAM_BATTING_2B	Doubles by batters (2B)	Positive Impact on Wins
TEAM_BATTING_3B	Triples by batters (3B)	Positive Impact on Wins
TEAM_BATTING_HR	Homeruns by batters (4B)	Positive Impact on Wins
TEAM_BATTING_HBP	Walks by batters	Positive Impact on Wins
TEAM_BATTING_BB	Batters hit by pitch (get a free base)	Positive Impact on Wins
TEAM_BATTING_SO	Strikeouts by batters	Negative Impact on Wins
TEAM_BASERUN_SB	Stolen bases	Positive Impact on Wins
TEAM_BASERUN_CS	Caught stealing	Negative Impact on Wins
TEAM_FIELDING_E	Errors	Negative Impact on Wins
TEAM_FIELDING_DP	Double Plays	Positive Impact on Wins
TEAM_PITCHING_BB	Walks allowed	Negative Impact on Wins
TEAM_PITCHING_H	Hits allowed	Negative Impact on Wins
TEAM_PITCHING_HR	Homeruns allowed	Negative Impact on Wins
TEAM_PITCHING_SO	Strikeouts by pitchers	Positive Impact on Wins

We can use the command read.csv to import the dataset and view the first six row with the command head().

```
##   INDEX TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
## 1     1          39           1445             194              39
## 2     2          70           1339             219              22
## 3     3          86           1377             232              35
## 4     4          70           1387             209              38
## 5     5          82           1297             186              27
## 6     6          75           1279             200              36
##   TEAM_BATTING_HR TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB
## 1              13             143             842              NA
## 2             190             685            1075              37
## 3             137             602             917              46
## 4              96             451             922              43
## 5             102             472             920              49
## 6              92             443             973             107
##   TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR
## 1              NA               NA            9364               84
## 2              28               NA            1347              191
## 3              27               NA            1377              137
## 4              30               NA            1396               97
## 5              39               NA            1297              102
## 6              59               NA            1279               92
##   TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E TEAM_FIELDING_DP
## 1              927             5456            1011               NA
## 2              689             1082             193              155
## 3              602              917             175              153
## 4              454              928             164              156
## 5              472              920             138              168
## 6              443              973             123              149
```

All the variables are numeric. The summary and describe function gives the univariate statistic of each variable. For each variable there are computation of minimun, maximun, mean, median, first and third quantiles. The describe function also include the standard deviation, the degree of skweness and the degree of kurtosis. For a quick univariate statistics of the datasets, the function summary is convenient.

Univariate Summary Statistics:

```
##   TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
##  Min.   :  0.00   Min.   : 891   Min.   : 69.0   Min.   :  0.00
##  1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00
##  Median : 82.00   Median :1454   Median :238.0   Median : 47.00
##  Mean   : 80.79   Mean   :1469   Mean   :241.2   Mean   : 55.25
##  3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00
##  Max.   :146.00   Max.   :2554   Max.   :458.0   Max.   :223.00
##
##  TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB
##  Min.   :  0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.0
##  1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0   1st Qu.: 66.0
##  Median :102.00   Median :512.0   Median : 750.0   Median :101.0
##  Mean   : 99.61   Mean   :501.6   Mean   : 735.6   Mean   :124.8
##  3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0   3rd Qu.:156.0  
##  Max.   :264.00   Max.   :878.0   Max.   :1399.0   Max.   :697.0  
##                                   NA's   :102      NA's   :131    
##  TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR
##  Min.   :  0.0   Min.   :29.00    Min.   : 1137   Min.   :  0.0  
##  1st Qu.: 38.0   1st Qu.:50.50    1st Qu.: 1419   1st Qu.: 50.0  
##  Median : 49.0   Median :58.00    Median : 1518   Median :107.0  
##  Mean   : 52.8   Mean   :59.36    Mean   : 1779   Mean   :105.7  
##  3rd Qu.: 62.0   3rd Qu.:67.00    3rd Qu.: 1682   3rd Qu.:150.0  
##  Max.   :201.0   Max.   :95.00    Max.   :30132   Max.   :343.0  
##  NA's   :772     NA's   :2085                                    
##  TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E  TEAM_FIELDING_DP
##  Min.   :   0.0   Min.   :    0.0   Min.   :  65.0   Min.   : 52.0  
##  1st Qu.: 476.0   1st Qu.:  615.0   1st Qu.: 127.0   1st Qu.:131.0  
##  Median : 536.5   Median :  813.5   Median : 159.0   Median :149.0  
##  Mean   : 553.0   Mean   :  817.7   Mean   : 246.5   Mean   :146.4  
##  3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2   3rd Qu.:164.0  
##  Max.   :3645.0   Max.   :19278.0   Max.   :1898.0   Max.   :228.0  
##                   NA's   :102                        NA's   :286    
```

All the variables are numeric There are missing values with variables TEAM_BATTING_SO, TEAM_BASERUN_SB, TEAM_BASERUN_CS, TEAM_BATTING_HBP, TEAM_PITCHING_SO, TEAM_FIELDING_DP.

In This train dataset, the target variable, TARGET_WINS, varies from 0 to 146.

The median and the mean are closed in values or in the same magnitude except TEAM_PITCHING_H where the mean is 200 time bigger than the median, TEAM_FIELDING_E where mean is also larger than median.

```
The inner structure of each variable can be obtained with the function
str in R.

## 'data.frame':    2276 obs. of  17 variables:
##  $ INDEX           : int  1 2 3 4 5 6 7 8 11 12 ...
##  $ TARGET_WINS     : int  39 70 86 70 82 75 80 85 86 76 ...
##  $ TEAM_BATTING_H  : int  1445 1339 1377 1387 1297 1279 1244 1273 1391
1271 ...
##  $ TEAM_BATTING_2B : int  194 219 232 209 186 200 179 171 197 213 ...
##  $ TEAM_BATTING_3B : int  39 22 35 38 27 36 54 37 40 18 ...
##  $ TEAM_BATTING_HR : int  13 190 137 96 102 92 122 115 114 96 ...
##  $ TEAM_BATTING_BB : int  143 685 602 451 472 443 525 456 447 441 ...
##  $ TEAM_BATTING_SO : int  842 1075 917 922 920 973 1062 1027 922 827 ...
##  $ TEAM_BASERUN_SB : int  NA 37 46 43 49 107 80 40 69 72 ...
##  $ TEAM_BASERUN_CS : int  NA 28 27 30 39 59 54 36 27 34 ...
##  $ TEAM_BATTING_HBP: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ TEAM_PITCHING_H : int  9364 1347 1377 1396 1297 1279 1244 1281 1391
1271 ...
##  $ TEAM_PITCHING_HR: int  84 191 137 97 102 92 122 116 114 96 ...
##  $ TEAM_PITCHING_BB: int  927 689 602 454 472 443 525 459 447 441 ...
##  $ TEAM_PITCHING_SO: int  5456 1082 917 928 920 973 1062 1033 922 827 ...
##  $ TEAM_FIELDING_E : int  1011 193 175 164 138 123 136 112 127 131 ...
##  $ TEAM_FIELDING_DP: int  NA 155 153 156 168 149 186 136 169 159 ...
```

The str function explains the structure of data frame. The data frame has 15 variables of type integer with 2276 observations

The histograms below allow the visualization of the distribution of each variable.

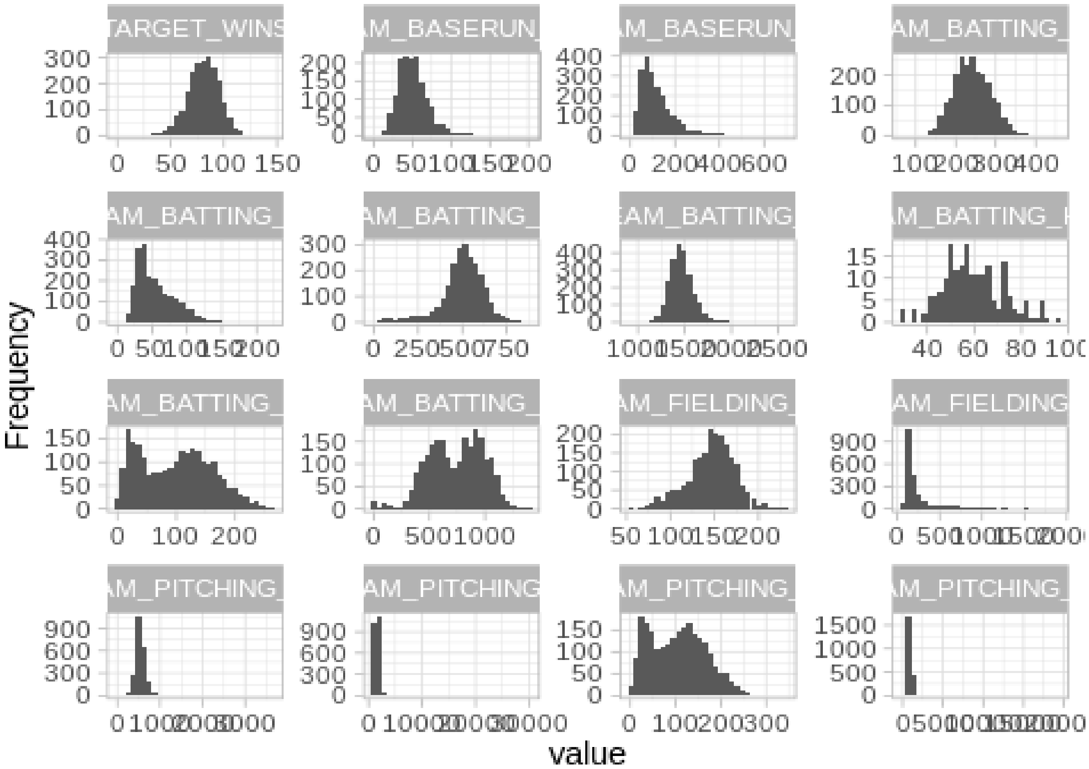


Fig. 1: Histograms showing distribution of each variable

The TARGET_WINS which is the target variable present a normal distribution. The variables TEAM_PITCHING_H, TEAM_PITCHING_BB, and TEAM_PITCHING_SO have high degrees of skweness and kurtosis. These variable need to be log transformed before introducting in a model.

There are three bimodal distributions TEAM_BATTING_HR, TEAM_BATING_SO, AND REAM_PITCHING_HR

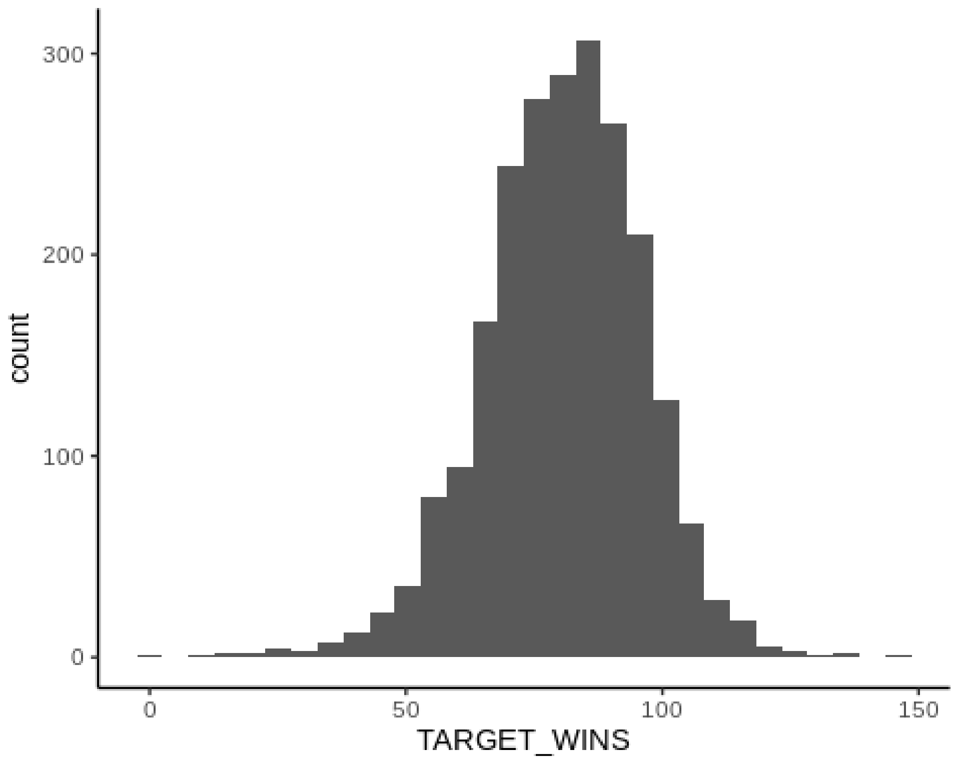


Fig. 2: Normal distribution of the target variable.

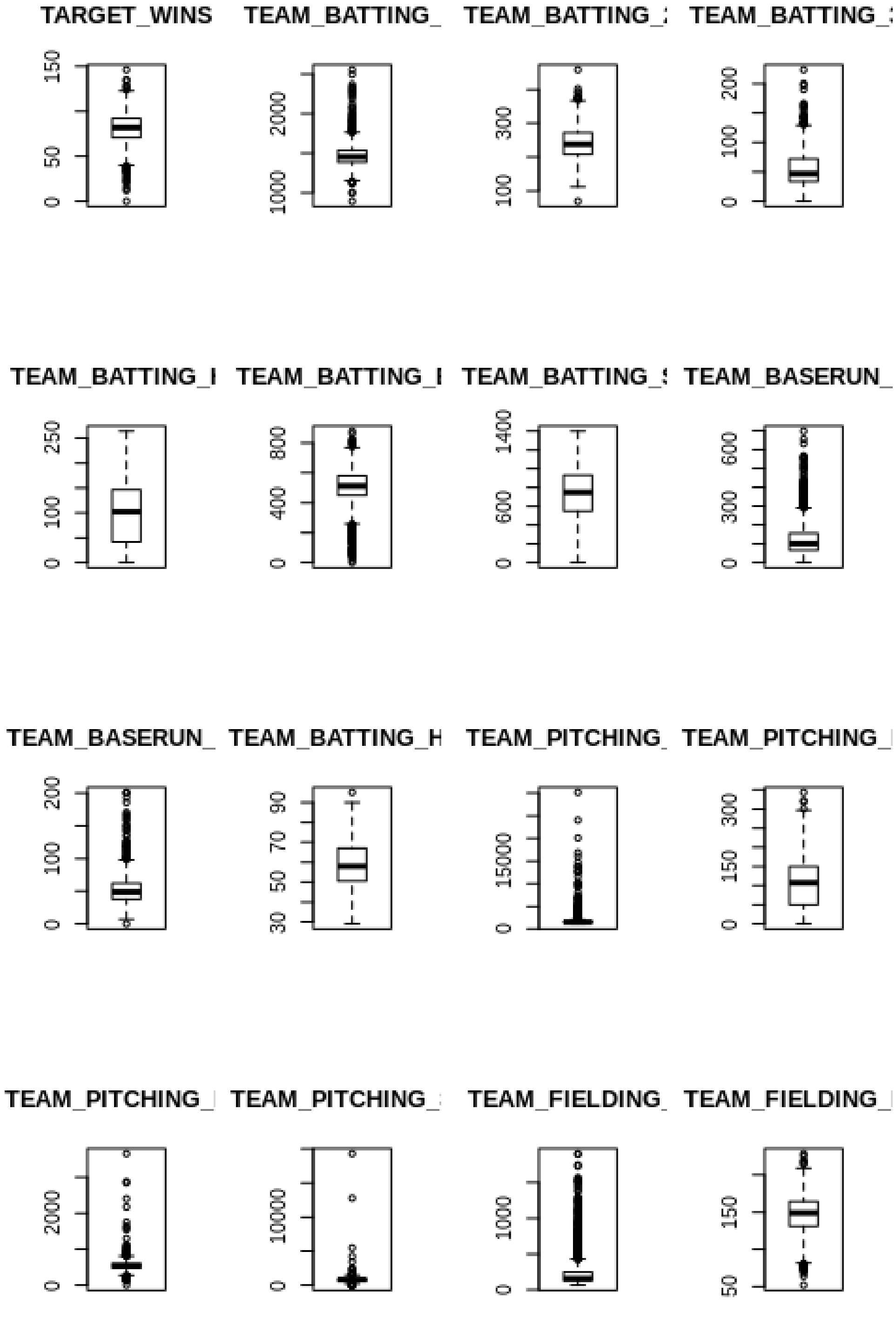


Fig.3: Boxplots for each variable

The boxplots of different variables add some visual information about the outliers. Some variable distributions are skewed by to much outlierS in one side as TEAM_FIELDING_E, TEAM_PITCHING_H, TEAM_BASERUN_CS, and TEAM_BATING_HR.

```
With the skewness of package e1071, we can find to what extend a variable is
skewed.

##      TARGET_WINS   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B
##       -0.3987232        1.5713335        0.2151018        1.1094652
##   TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB
##        0.1860421       -1.0257599               NA               NA
##   TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H TEAM_PITCHING_HR
##               NA               NA       10.3295111        0.2877877
## TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E TEAM_FIELDING_DP
##        6.7438995               NA        2.9904656               NA
```

The aggr function in the VIM package plots and calculates the amount of missing values in each variable. The dplyr function is useful for wrangling data into aggregate summaries and is used to find the pattern of missing data related to the classes.

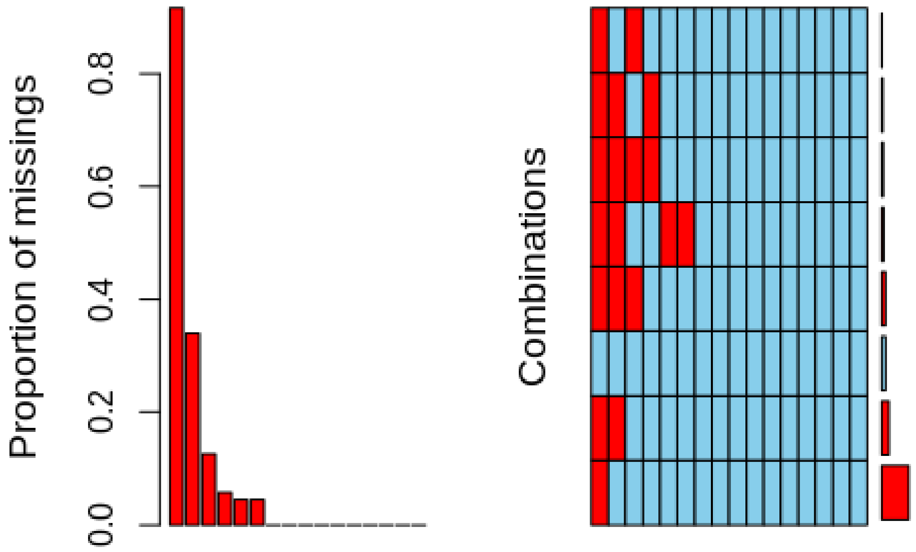


Fig. 4: Graph of Missings Data

```
##
##  Variables sorted by number of missings:
##          Variable      Rate
##  TEAM_BATTING_HBP 0.91608084
##   TEAM_BASERUN_CS 0.33919156
##  TEAM_FIELDING_DP 0.12565905
##   TEAM_BASERUN_SB 0.05755712
##   TEAM_BATTING_SO 0.04481547
##  TEAM_PITCHING_SO 0.04481547
##       TARGET_WINS 0.00000000
##    TEAM_BATTING_H 0.00000000
##   TEAM_BATTING_2B 0.00000000
##   TEAM_BATTING_3B 0.00000000
##   TEAM_BATTING_HR 0.00000000
##   TEAM_BATTING_BB 0.00000000
##   TEAM_PITCHING_H 0.00000000
##  TEAM_PITCHING_HR 0.00000000
##  TEAM_PITCHING_BB 0.00000000
##   TEAM_FIELDING_E 0.00000000
```

TEAM_BATTING_HBP and TEAM_BASERUN_CS have respectively 91.6% and 34% fo missing values in their respective column. Including those variable in the model imply an imputation of massive data in the model. We will exclude those variables from the model.

The correlations between variables in our training dataset are below.

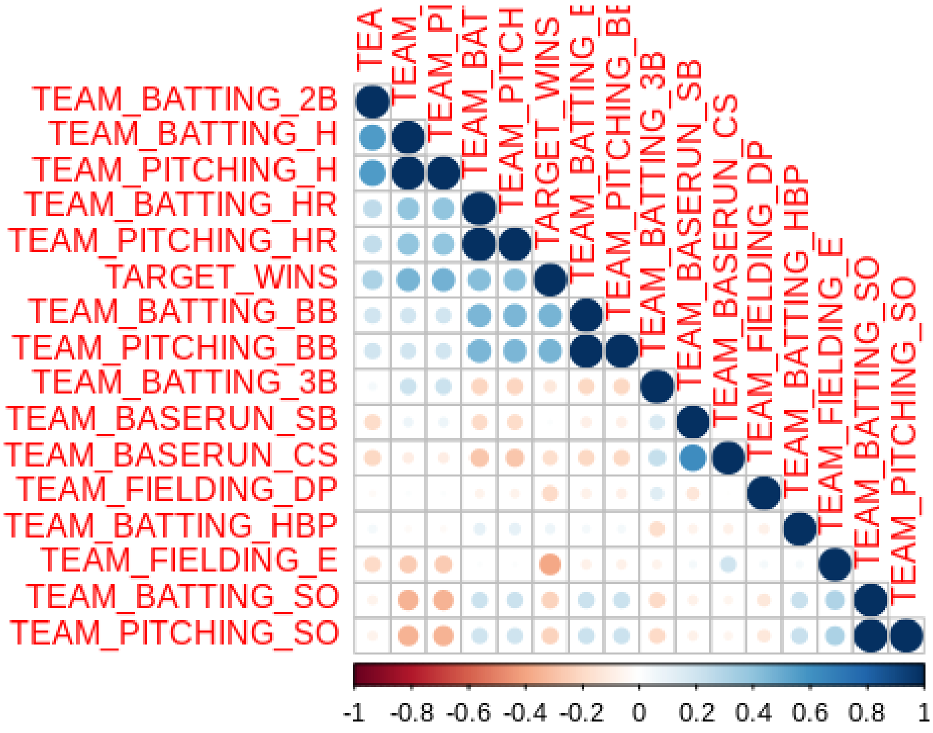


Fig. 5: Correlation graph

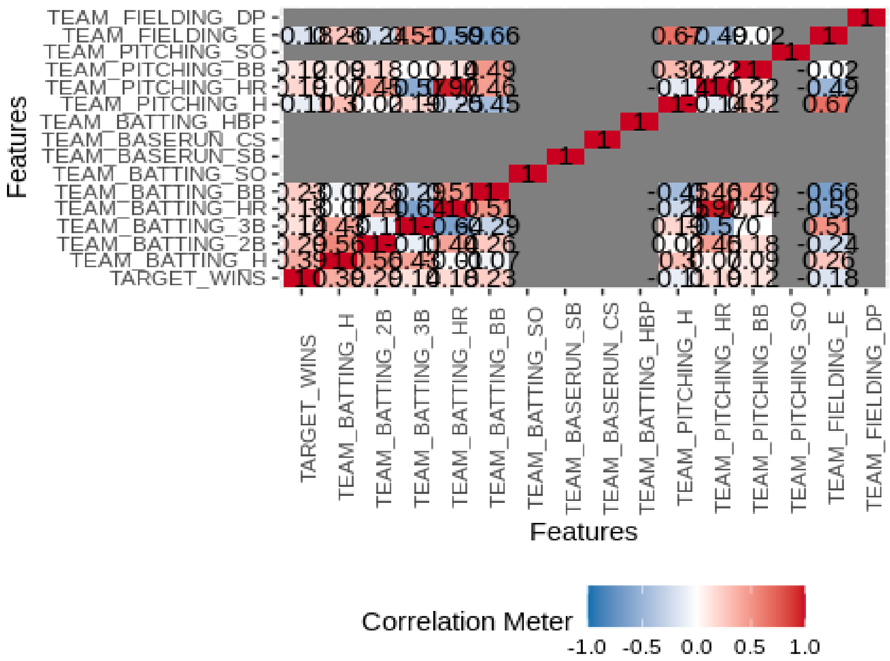


Fig. 6: Another correlation graph

There is no strong correlation between the target variable with other predictors.

Divers Correlations with TARGET_WINS

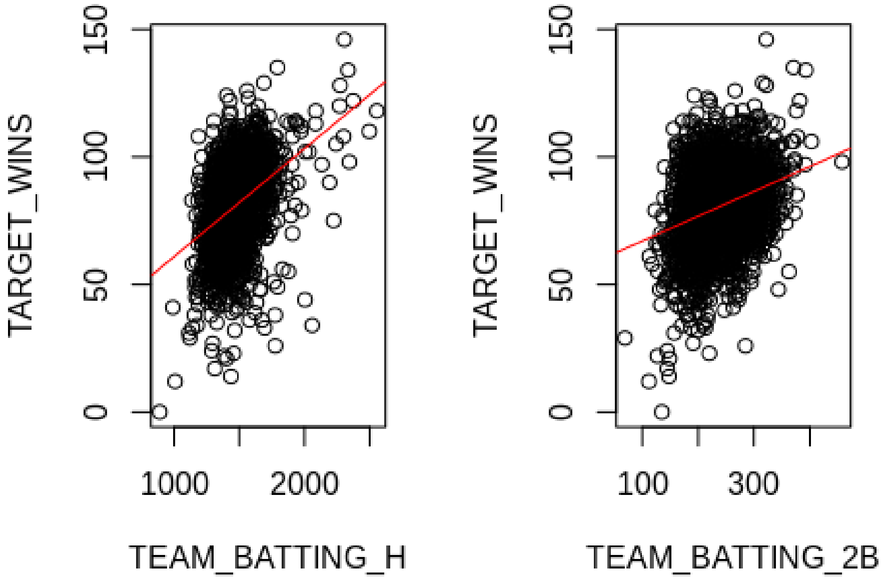


Fig. 7: Relationship with the target variable

High correlated predictors

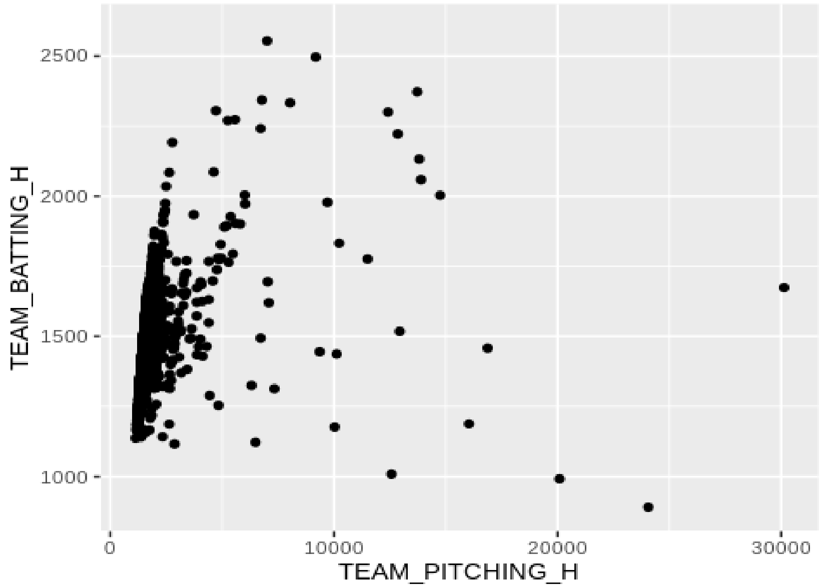


Fig. 8:

There exist a trend in the relationship between TEAM_BATTING_H and TEAM_PITCHING_H

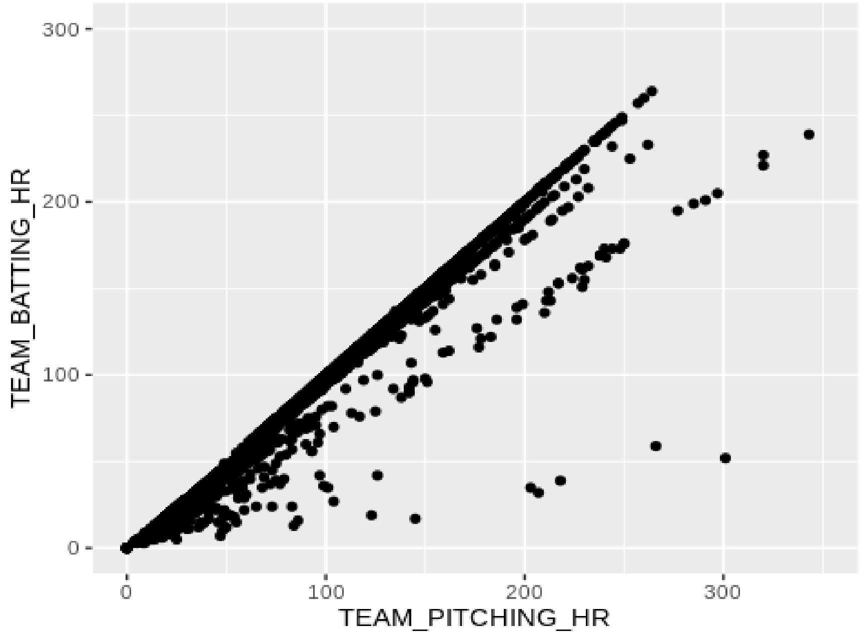


Fig. 9:

The relation between TEAM_BATTING_HR and TEAM_PITCHING_H is strong enough even though there are multiple layers of linearities.

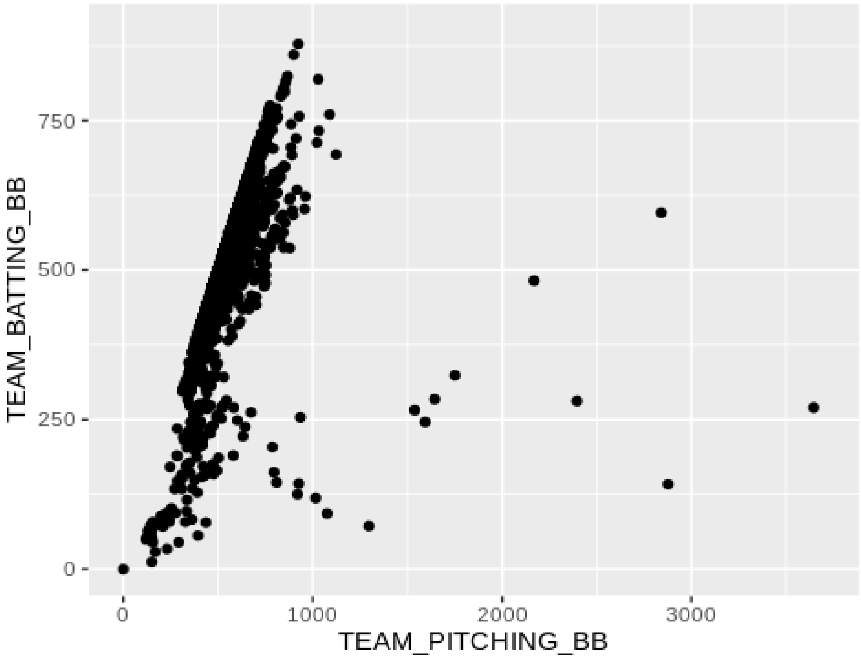


Fig. 10:

TEAM_BATTING_BB and TEAM_PITCHING_BB could be collinear if we remove some outliers that leverage the relationship.

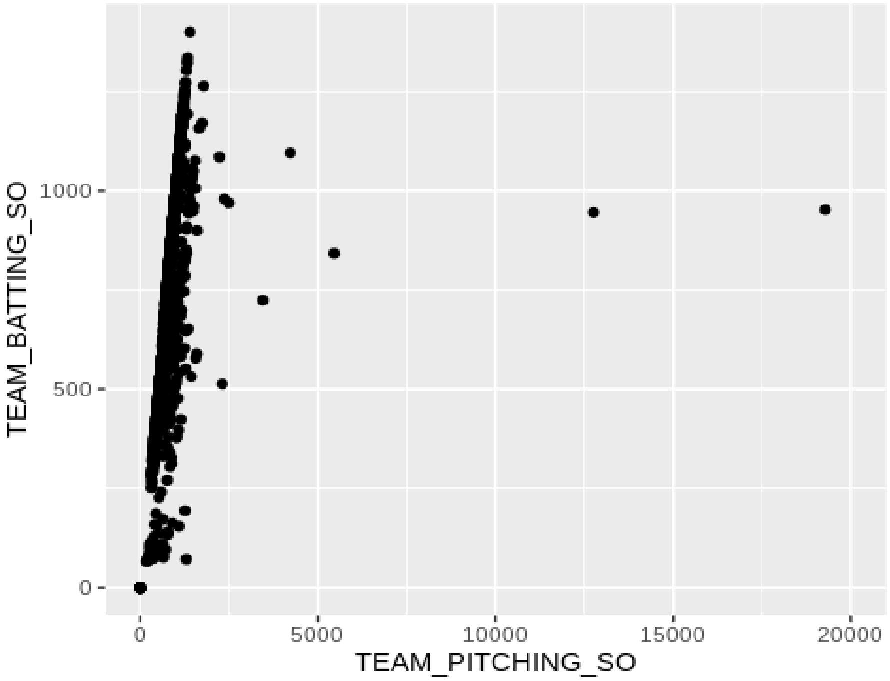


Fig. 11:

TEAM_BATTING_SO AND TEAM_PITCHING_SO are colinear at some levels.

DATA PREPARATION

Remove the two variables with lot of missing data

Imputing the median in place of missing data

```
##            INDEX      TARGET_WINS   TEAM_BATTING_H  TEAM_BATTING_2B 
##                0                0                0                0 
##  TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO 
##                0                0                0                0 
##  TEAM_BASERUN_SB  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H 
##                0                0                0                0 
## TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E 
##                0                0                0                0 
## TEAM_FIELDING_DP 
##                0
```

Splitting into train test dataset

Transforming the skewed variables

Look for lambda transformation

```
## Box-Cox Transformation
## 
## 2276 data points used to estimate Lambda
## 
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    65.0   127.0   159.0   246.5   249.2  1898.0 
## 
## Largest/Smallest: 29.2 
## Sample Skewness: 2.99 
## 
## Estimated Lambda: -0.9
```

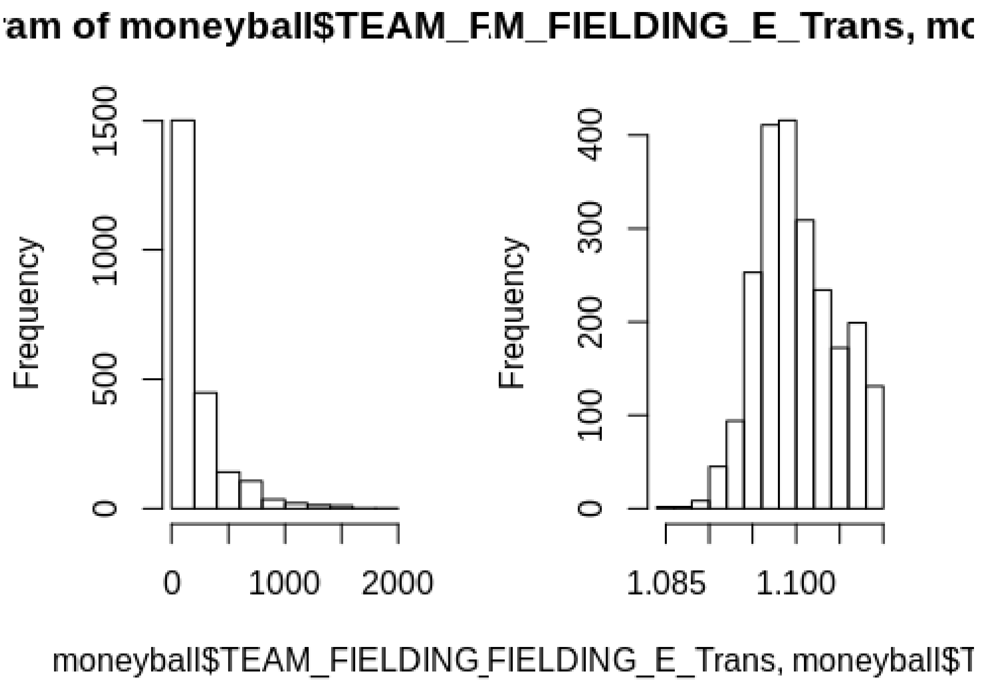


Fig12: Histograms of predictor TEAM+FIELDING before and after the transformations

```
## Created from 2276 samples and 17 variables
##
## Pre-processing:
##   - Box-Cox transformation (7)
##   - centered (17)
##   - ignored (0)
##   - principal component signal extraction (17)
##   - scaled (17)
##
## Lambda estimates for Box-Cox transformation:
## 0.7, -1.3, 0.6, 0.4, -2, -0.9, 1.8
## PCA needed 11 components to capture 95 percent of the variance
```

```
# Apply the transformations:
```

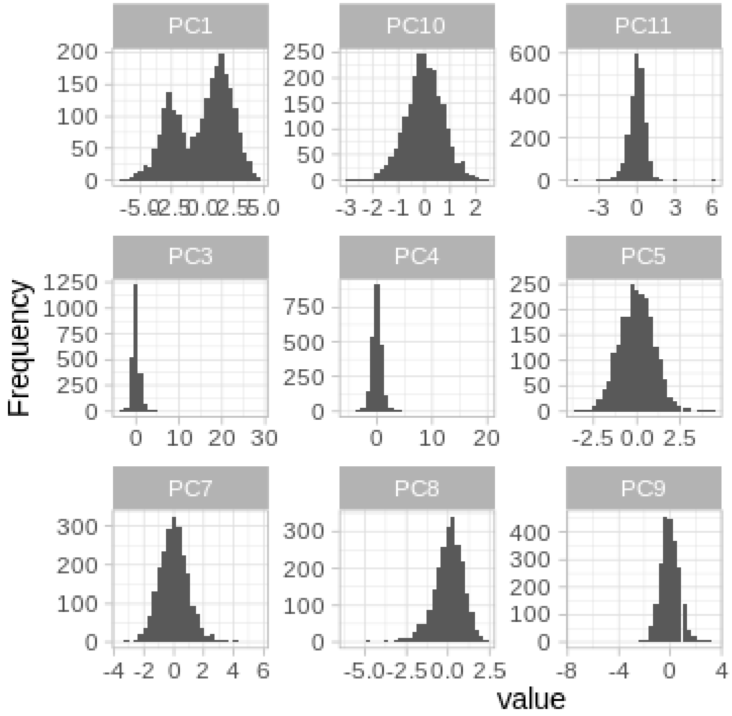


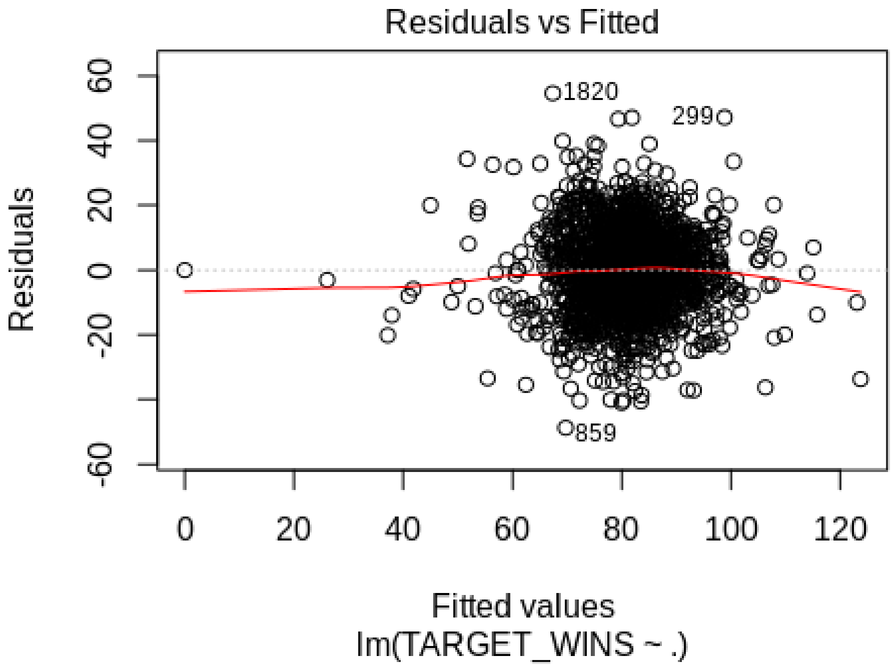
Fig 13: Histograms of Principal components

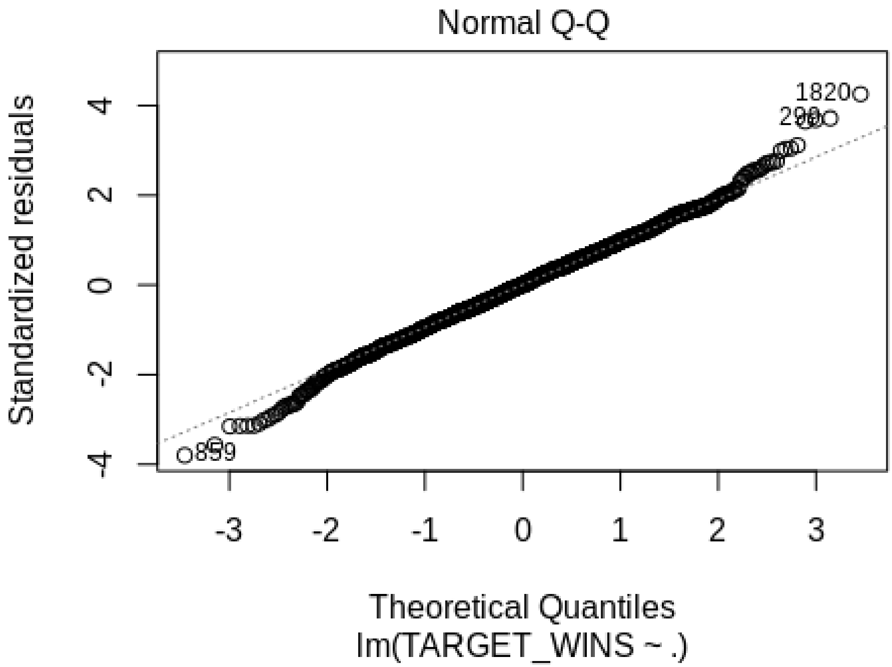
```
transformed[1:6,1:5]

##           PC1        PC2         PC3        PC4         PC5
## 1 -1.6077922 -1.1719114  5.64426824  7.4100674  0.09615263
## 2  3.0691749 -1.0499726  0.53511064  0.5932772 -1.90541052
## 3  1.7792841 -0.5512956 -0.07011525  0.2485807 -1.57886869
## 4  0.7338135 -1.5622378 -0.85211911  0.9274248 -1.40218521
## 5  1.3572663 -2.1587298 -0.61502486  0.1515135 -1.65495497
## 6  0.9350596 -2.5850492 -0.25872530 -0.5360716 -1.00937426

#colSums(is.na(moneyballp))
```

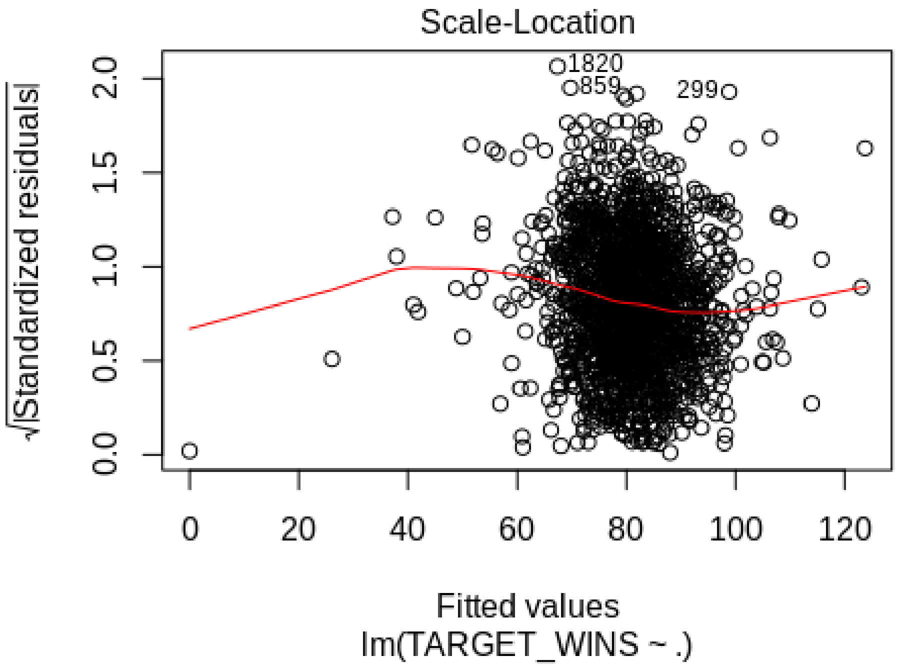
BUILD MODELS

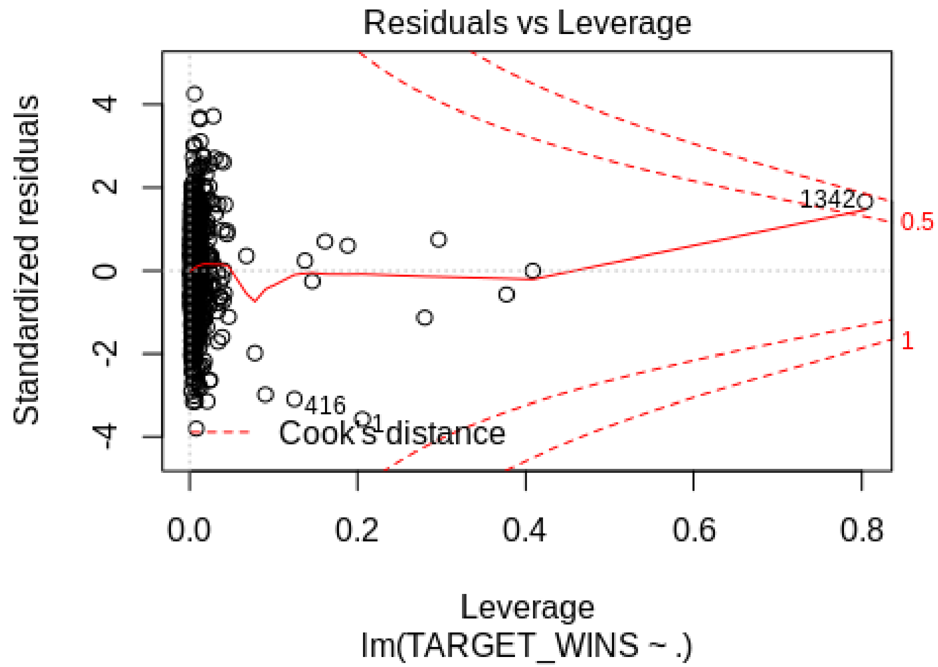
This first model takes all selected predictors into account.

```
lm01 <- lm(TARGET_WINS~., moneyball_train1)
summary(lm01)

## 
## Call:
## lm(formula = TARGET_WINS ~ ., data = moneyball_train1)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -48.734  -8.124   0.001   8.288  54.604
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       26.4877901  5.8872286   4.499 7.26e-06 ***
## TEAM_BATTING_H     0.0448682  0.0040650  11.038  < 2e-16 ***
## TEAM_BATTING_2B   -0.0307027  0.0101105  -3.037  0.00243 **
## TEAM_BATTING_3B    0.0968992  0.0181473   5.340 1.05e-07 ***
## TEAM_BATTING_HR    0.0467460  0.0313119   1.493  0.13563
## TEAM_BATTING_BB    0.0190663  0.0069012   2.763  0.00579 **
## TEAM_BATTING_SO   -0.0136139  0.0032407  -4.201 2.79e-05 ***
## TEAM_BASERUN_SB    0.0278951  0.0048189   5.789 8.35e-09 ***
## TEAM_PITCHING_H   -0.0002383  0.0004089  -0.583  0.56017
## TEAM_PITCHING_HR   0.0396934  0.0277145   1.432  0.15225
## TEAM_PITCHING_BB  -0.0058496  0.0050952  -1.148  0.25110
## TEAM_PITCHING_SO   0.0072298  0.0016350   4.422 1.04e-05 ***
## TEAM_FIELDING_E   -0.0200219  0.0026864  -7.453 1.41e-13 ***
## TEAM_FIELDING_DP  -0.1258433  0.0140361  -8.966  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.87 on 1806 degrees of freedom
## Multiple R-squared:  0.3195, Adjusted R-squared:  0.3146
## F-statistic: 65.22 on 13 and 1806 DF,  p-value: < 2.2e-16

plot(lm01)
```


Residuals vs Fitted
Residuals
60
40
20
0
-20
-60
1820
299
859
0
20
40
60
80
100
120
Fitted values
lm(TARGET_WINS ~ .)


Normal Q-Q
Standardized residuals
4
2
0
-2
-4
1820
299
859
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(TARGET_WINS ~ .)





We remove the predictor with the highest p-value

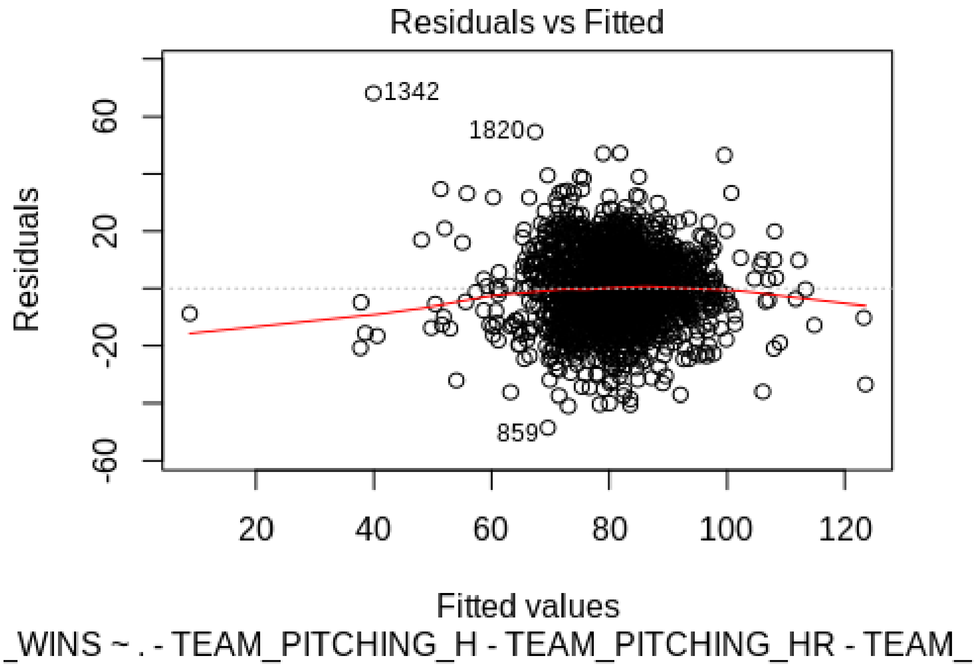
```
lm02 <- lm(TARGET_WINS~.-TEAM_BATTING_SO, moneyball_train1)
summary(lm02)
```

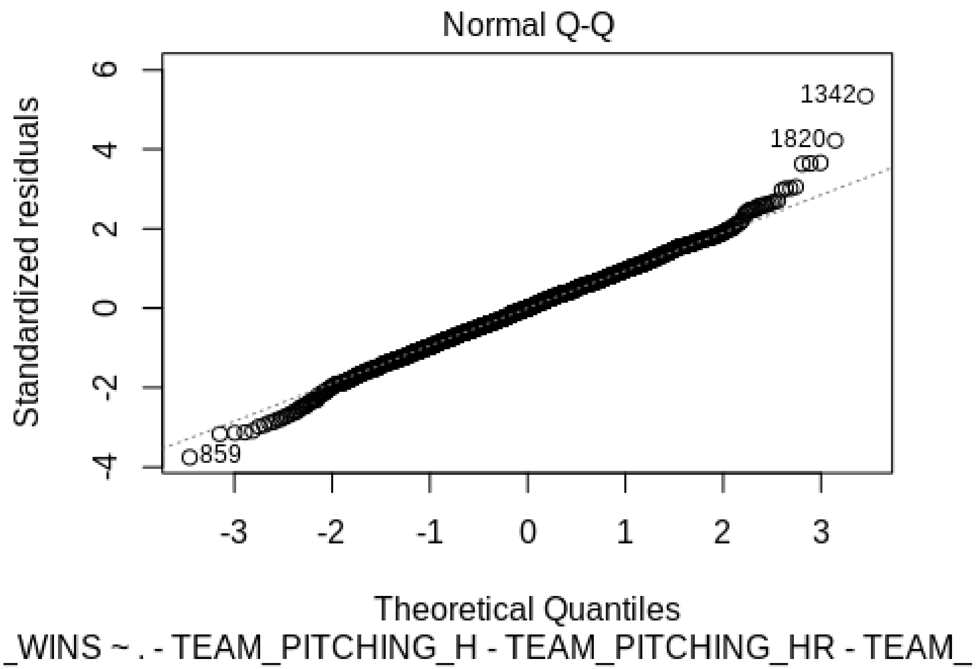
```
##
## Call:
## lm(formula = TARGET_WINS ~ . - TEAM_BATTING_SO, data = moneyball_train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.680  -8.466  -0.020   8.395  52.873
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      12.4297738  4.8658784   2.554 0.010716 *
## TEAM_BATTING_H    0.0511463  0.0037976  13.468  < 2e-16 ***
## TEAM_BATTING_2B  -0.0360255  0.0100769  -3.575 0.000359 ***
## TEAM_BATTING_3B   0.1036119  0.0181599   5.706 1.35e-08 ***
## TEAM_BATTING_HR   0.0400851  0.0314154   1.276 0.202131
## TEAM_BATTING_BB   0.0116223  0.0067004   1.735 0.082989 .
## TEAM_BASERUN_SB   0.0215647  0.0045982   4.690 2.94e-06 ***
## TEAM_PITCHING_H  -0.0003091  0.0004104  -0.753 0.451437
## TEAM_PITCHING_HR  0.0216824  0.0275067   0.788 0.430648
## TEAM_PITCHING_BB  0.0031008  0.0046497   0.667 0.504935
## TEAM_PITCHING_SO  0.0030405  0.0013017   2.336 0.019608 *
## TEAM_FIELDING_E  -0.0187112  0.0026805  -6.981 4.12e-12 ***
## TEAM_FIELDING_DP -0.1186085  0.0139941  -8.476  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.93 on 1807 degrees of freedom
## Multiple R-squared:  0.3128, Adjusted R-squared:  0.3083
## F-statistic: 68.55 on 12 and 1807 DF,  p-value: < 2.2e-16
```

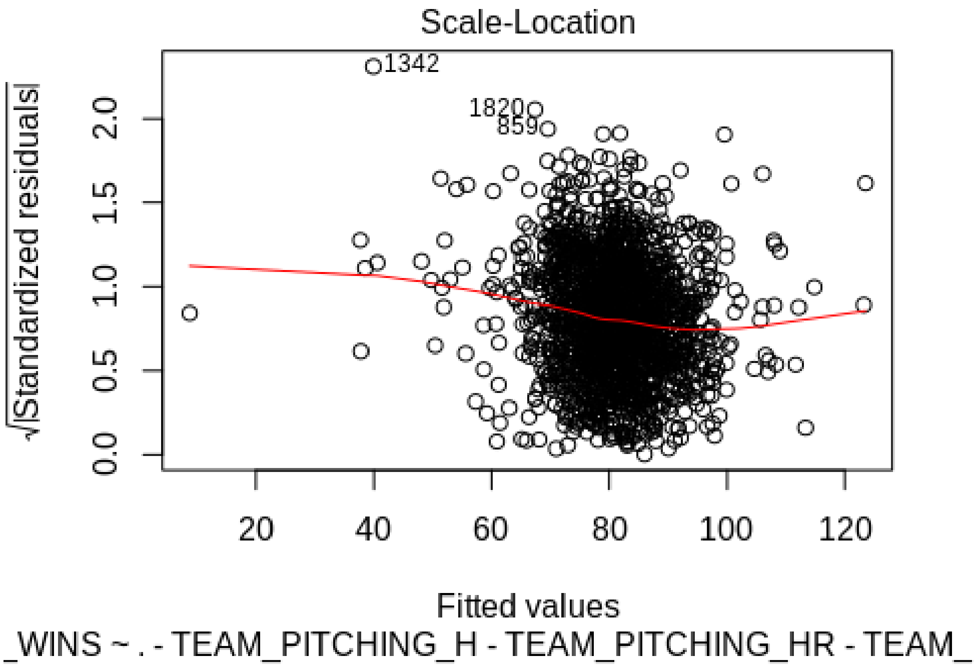
```
lm11 <- lm(TARGET_WINS~.-TEAM_PITCHING_H-TEAM_PITCHING_HR-TEAM_PITCHING_BB-
TEAM_PITCHING_SO, moneyball_train1)
summary(lm11)
```

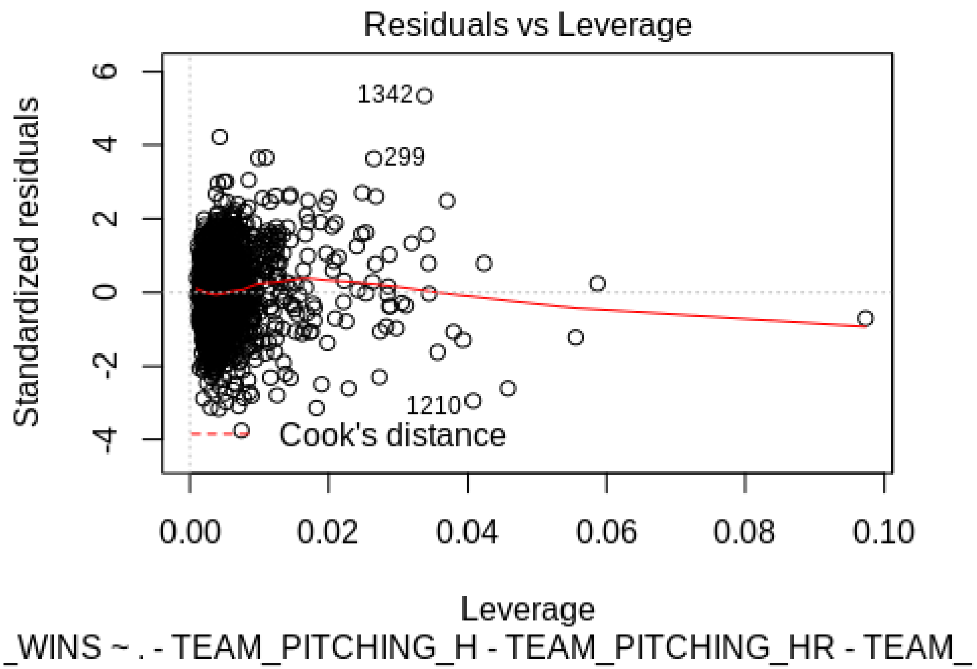
```
##
## Call:
## lm(formula = TARGET_WINS ~ . - TEAM_PITCHING_H - TEAM_PITCHING_HR -
##     TEAM_PITCHING_BB - TEAM_PITCHING_SO, data = moneyball_train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.583  -8.284  -0.017   8.285  68.063
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      25.415618   5.870164   4.330 1.58e-05 ***
## TEAM_BATTING_H    0.043405   0.004022  10.793  < 2e-16 ***
## TEAM_BATTING_2B  -0.023340   0.010099  -2.311  0.02094 *
## TEAM_BATTING_3B   0.100211   0.017733   5.651 1.85e-08 ***
## TEAM_BATTING_HR   0.080643   0.010851   7.432 1.64e-13 ***
## TEAM_BATTING_BB   0.012287   0.003774   3.255  0.00115 **
## TEAM_BATTING_SO  -0.004257   0.002580  -1.650  0.09909 .
## TEAM_BASERUN_SB   0.026545   0.004736   5.605 2.40e-08 ***
## TEAM_FIELDING_E  -0.017943   0.002203  -8.145 7.00e-16 ***
## TEAM_FIELDING_DP -0.122708   0.014141  -8.678  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.98 on 1810 degrees of freedom
## Multiple R-squared:  0.3062, Adjusted R-squared:  0.3028
## F-statistic: 88.77 on 9 and 1810 DF,  p-value: < 2.2e-16
```

```
plot(lm11)
```


Residuals vs Fitted
Residuals
Fitted values
_WINS ~ . - TEAM_PITCHING_H - TEAM_PITCHING_HR - TEAM_
60
20
-20
-60
20
40
60
80
100
120
1342
1820
859


Normal Q-Q
Standardized residuals
Theoretical Quantiles
_WINS ~ . - TEAM_PITCHING_H - TEAM_PITCHING_HR - TEAM_
6
4
2
0
-2
-4
-3
-2
-1
0
1
2
3
1342
1820
859





```r
lm2 <- lm(TARGET_WINS~TEAM_BATTING_2B+TEAM_BATTING_H+TEAM_PITCHING_H+
            TEAM_BATTING_HR+TEAM_PITCHING_HR+
            TEAM_PITCHING_BB+TEAM_FIELDING_E , moneyball)
summary(lm2)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_H +
##     TEAM_PITCHING_H + TEAM_BATTING_HR + TEAM_PITCHING_HR +
TEAM_PITCHING_BB +
##     TEAM_FIELDING_E, data = moneyball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.298  -8.868   0.110   8.799  51.667
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.3576809  3.3445960   0.406 0.684830
## TEAM_BATTING_2B  -0.0334941  0.0090405  -3.705 0.000217 ***
## TEAM_BATTING_H    0.0585216  0.0028059  20.856  < 2e-16 ***
## TEAM_PITCHING_H  -0.0018772  0.0003147  -5.965 2.83e-09 ***
## TEAM_BATTING_HR   0.0164810  0.0240922   0.684 0.493995
## TEAM_PITCHING_HR -0.0047054  0.0226647  -0.208 0.835554
## TEAM_PITCHING_BB  0.0128687  0.0020225   6.363 2.39e-10 ***
## TEAM_FIELDING_E  -0.0137590  0.0022656  -6.073 1.47e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.59 on 2268 degrees of freedom
## Multiple R-squared:  0.2582, Adjusted R-squared:  0.2559
## F-statistic: 112.8 on 7 and 2268 DF,  p-value: < 2.2e-16
```

```
lm3 <- lm(TARGET_WINS~TEAM_BATTING_2B+TEAM_BATTING_H+
            TEAM_BATTING_HR+TEAM_BATTING_SO+
            TEAM_BATTING_BB, moneyball)
summary(lm3)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_H +
##     TEAM_BATTING_HR + TEAM_BATTING_SO + TEAM_BATTING_BB, data = moneyball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.904  -8.595   0.573   8.982  53.284
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -9.937644   5.012318  -1.983  0.04753 *
## TEAM_BATTING_2B -0.022361   0.009297  -2.405  0.01625 *
## TEAM_BATTING_H   0.052027   0.003378  15.404  < 2e-16 ***
## TEAM_BATTING_HR  0.025820   0.008622   2.995  0.00278 **
## TEAM_BATTING_SO  0.002612   0.002260   1.156  0.24777
## TEAM_BATTING_BB  0.029388   0.002772  10.601  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.58 on 2168 degrees of freedom
##   (102 observations deleted due to missingness)
## Multiple R-squared:  0.2419, Adjusted R-squared:  0.2401
## F-statistic: 138.4 on 5 and 2168 DF,  p-value: < 2.2e-16

lm3 <-
lm(TARGET_WINS~TEAM_BATTING_2B+TEAM_PITCHING_H+TEAM_PITCHING_HR+TEAM_PITCHING
_SO+TEAM_PITCHING_BB, moneyball)
summary(lm3)

##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_2B + TEAM_PITCHING_H +
##     TEAM_PITCHING_HR + TEAM_PITCHING_SO + TEAM_PITCHING_BB, data =
moneyball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.118  -9.519   0.245   9.378  67.184
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      54.6413858  1.8825343  29.025  < 2e-16 ***
## TEAM_BATTING_2B   0.0827264  0.0075317  10.984  < 2e-16 ***
## TEAM_PITCHING_H  -0.0012829  0.0002392  -5.363 9.05e-08 ***
## TEAM_PITCHING_HR  0.0219313  0.0060492   3.625 0.000295 ***
## TEAM_PITCHING_SO -0.0048607  0.0006606  -7.357 2.65e-13 ***
## TEAM_PITCHING_BB  0.0176132  0.0022130   7.959 2.77e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.47 on 2168 degrees of freedom
##   (102 observations deleted due to missingness)
## Multiple R-squared:  0.1385, Adjusted R-squared:  0.1365
## F-statistic: 69.71 on 5 and 2168 DF,  p-value: < 2.2e-16

lm2 <- lm(TARGET_WINS~TEAM_BATTING_2B+TEAM_BATTING_H+TEAM_PITCHING_H+

TEAM_BATTING_HR+TEAM_PITCHING_HR+TEAM_BATTING_SO+TEAM_PITCHING_SO+
            TEAM_BATTING_BB+TEAM_PITCHING_BB, moneyball)
summary(lm2)

##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_H +
##     TEAM_PITCHING_H + TEAM_BATTING_HR + TEAM_PITCHING_HR + TEAM_BATTING_SO
+
##     TEAM_PITCHING_SO + TEAM_BATTING_BB + TEAM_PITCHING_BB, data =
moneyball)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -51.641  -8.660   0.346   9.026  49.760 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      -4.7512713  5.0667223  -0.938  0.34848    
## TEAM_BATTING_2B  -0.0259231  0.0092470  -2.803  0.00510 ** 
## TEAM_BATTING_H    0.0562882  0.0034454  16.337  < 2e-16 ***
## TEAM_PITCHING_H  -0.0026602  0.0003322  -8.007  1.9e-15 ***
## TEAM_BATTING_HR   0.0329398  0.0270819   1.216  0.22400    
## TEAM_PITCHING_HR  0.0065137  0.0246717   0.264  0.79179    
## TEAM_BATTING_SO  -0.0041868  0.0025209  -1.661  0.09689 .  
## TEAM_PITCHING_SO  0.0027962  0.0009324   2.999  0.00274 ** 
## TEAM_BATTING_BB   0.0149152  0.0057096   2.612  0.00906 ** 
## TEAM_PITCHING_BB  0.0049287  0.0041793   1.179  0.23841    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 13.35 on 2164 degrees of freedom
##   (102 observations deleted due to missingness)
## Multiple R-squared:  0.2685, Adjusted R-squared:  0.2655 
## F-statistic: 88.26 on 9 and 2164 DF,  p-value: < 2.2e-16
```

Tuning Linear Model

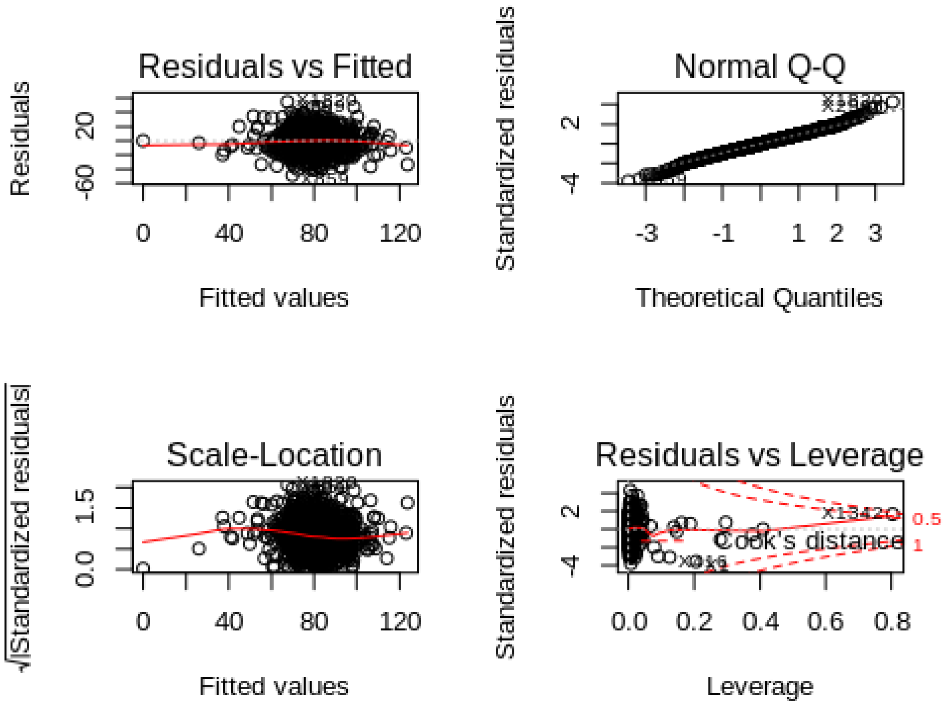
Model using tuning parameters

```
## Linear Regression 
## 
## 1820 samples
##   13 predictor
## 
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times) 
## Summary of sample sizes: 1638, 1637, 1639, 1638, 1638, 1639, ... 
## Resampling results:
## 
##   RMSE      Rsquared   MAE     
##   13.00564  0.3011865  10.20265
## 
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
summary(lmg)
```

```
## 
## Call:
## lm(formula = .outcome ~ ., data = dat)
## 
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -48.734  -8.124   0.001   8.288  54.604 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      26.4877901  5.8872286   4.499 7.26e-06 ***
## TEAM_BATTING_H    0.0448682  0.0040650  11.038  < 2e-16 ***
## TEAM_BATTING_2B  -0.0307027  0.0101105  -3.037  0.00243 ** 
## TEAM_BATTING_3B   0.0968992  0.0181473   5.340 1.05e-07 ***
## TEAM_BATTING_HR   0.0467460  0.0313119   1.493  0.13563    
## TEAM_BATTING_BB   0.0190663  0.0069012   2.763  0.00579 ** 
## TEAM_BATTING_SO  -0.0136139  0.0032407  -4.201 2.79e-05 ***
## TEAM_BASERUN_SB   0.0278951  0.0048189   5.789 8.35e-09 ***
## TEAM_PITCHING_H  -0.0002383  0.0004089  -0.583  0.56017    
## TEAM_PITCHING_HR  0.0396934  0.0277145   1.432  0.15225    
## TEAM_PITCHING_BB -0.0058496  0.0050952  -1.148  0.25110    
## TEAM_PITCHING_SO  0.0072298  0.0016350   4.422 1.04e-05 ***
## TEAM_FIELDING_E  -0.0200219  0.0026864  -7.453 1.41e-13 ***
## TEAM_FIELDING_DP -0.1258433  0.0140361  -8.966  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 12.87 on 1806 degrees of freedom
## Multiple R-squared:  0.3195, Adjusted R-squared:  0.3146 
## F-statistic: 65.22 on 13 and 1806 DF,  p-value: < 2.2e-16
```



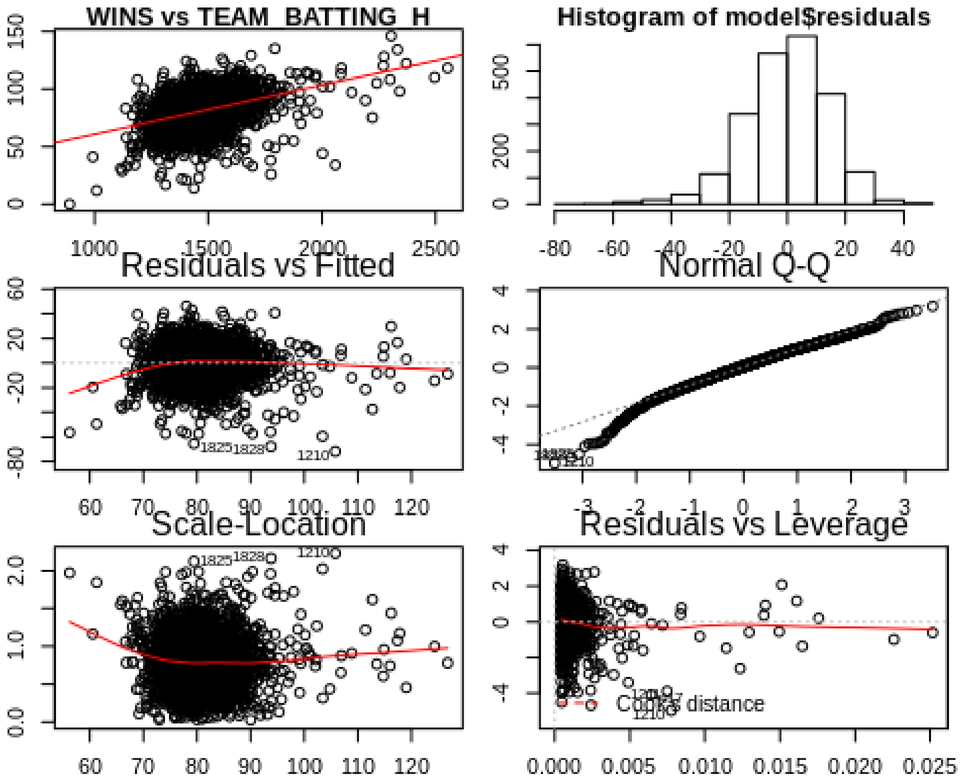
Model for pca

```
s #
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.068  -8.866   0.519   9.114  58.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  80.8630     0.3221 251.059  < 2e-16 ***
## PC1          -0.4972     0.1432  -3.472 0.000528 ***
## PC2          -3.7905     0.2076 -18.254  < 2e-16 ***
## PC3          -0.2585     0.2417  -1.070 0.284944
## PC4           1.9969     0.2918   6.844 1.05e-11 ***
## PC5           4.1191     0.3378  12.194  < 2e-16 ***
## PC6          -0.9238     0.4257  -2.170 0.030127 *
## PC7           0.9910     0.4727   2.096 0.036182 *
## PC8          -2.6243     0.5772  -4.547 5.81e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.74 on 1813 degrees of freedom
## Multiple R-squared:  0.2368, Adjusted R-squared:  0.2335
## F-statistic: 70.33 on 8 and 1813 DF,  p-value: < 2.2e-16
```

Foward Selection

Impact of each predictor on the outcome

The following series of plots show the type of relationship between the target variable and the predictors. For each group of plots, there are a scatter plot TARGET_WINS against the predictors, the histogram of residuals, The scatter plot of residuals against the fitted values, the normal quantile, the scale location, and the residuals against the leverage.



Residual vs Fitted of TARGET_WINS vs BATTING_HR shows heteroscedascity

Below is the adjusted R Squared of different model of TARGET_WINS against each predictor.

```
##            INDEX      TARGET_WINS   TEAM_BATTING_H  TEAM_BATTING_2B
##     3.814687e-06     1.000000e+00     1.507669e-01     8.317792e-02
##  TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO
##     1.990635e-02     3.060384e-02     5.366812e-02     4.958767e-04
##  TEAM_BASERUN_SB  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H
##     1.484661e-02    -1.849260e-04    -1.668419e-04     1.165172e-02
## TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
##     3.530215e-02     1.498634e-02     5.308364e-03     3.072081e-02
## TEAM_FIELDING_DP
##     4.658299e-04
```

Those values are very small. TARGET_WINS does not have a solid relationship with any those predictor. One predictor cannot explain significantly the target variable; therefor, multiple linear regression must be study.

Foward selection

We add one predictor at the time and observe the change in adjusted r_squared. If the r_squared increases, we keep the predictor, otherwise we remove that predictor.

```
## [1] 0.307881
```

```
summary(foward.selection.model)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
##     TEAM_BATTING_BB + TEAM_PITCHING_HR + TEAM_FIELDING_E + TEAM_BATTING_3B
+
##     TEAM_BASERUN_SB + TEAM_PITCHING_H + TEAM_PITCHING_SO +
TEAM_FIELDING_DP,
##     data = moneyball_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.016  -8.545   0.080   8.434  55.883
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.1840241  3.9184730   2.854  0.00435 **
## TEAM_BATTING_H    0.0540533  0.0033680  16.049  < 2e-16 ***
## TEAM_BATTING_2B  -0.0266846  0.0089866  -2.969  0.00302 **
## TEAM_BATTING_BB   0.0139419  0.0033243   4.194 2.85e-05 ***
## TEAM_PITCHING_HR  0.0416183  0.0069515   5.987 2.48e-09 ***
## TEAM_FIELDING_E  -0.0187712  0.0023813  -7.883 4.93e-15 ***
## TEAM_BATTING_3B   0.0663052  0.0159486   4.157 3.34e-05 ***
## TEAM_BASERUN_SB   0.0206002  0.0040528   5.083 4.02e-07 ***
## TEAM_PITCHING_H  -0.0006610  0.0003127  -2.114  0.03462 *
## TEAM_PITCHING_SO  0.0020188  0.0006208   3.252  0.00116 **
## TEAM_FIELDING_DP -0.1146370  0.0128170  -8.944  < 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:  0.3109, Adjusted R-squared:  0.3079
## F-statistic: 102.2 on 10 and 2265 DF,  p-value: < 2.2e-16
```

SELECT MODELS

We study many multiple linear regression models, we compare their Adjusted R Squared, RMSE, and RME. We split the data in training and testing set. These metrics result from the testing set of each linear model.

Here are the values of different metrics.

	R Squared	RMSE	MAE
Model1(Transformed)	0.22	13.74	10.59
Model2(Forward Selection)	0.36	12.88	10.21
Model3	0.21	14.59	11.24
Model4(Tuning Parameters)	0.28	13.44	10.30

The forward selection appears to be the best model. This model is significant since it p value is very low.

References

Applied Predictive Modeling Max Kuhn Kjell Jonhson

Appendix 1 Code

Code 1

https://github.com/AlainKuiete/DATA621/blob/master/DATA621Homework1.Rmd

Code 2

https://github.com/AlainKuiete/DATA621/blob/master/Assingment1.Rmd

Appendix 2 Predicted Values

https://github.com/AlainKuiete/DATA621/blob/master/moneyball_predict