Source : https://techcrunch.com/2016/04/02/moneyball-is-dead-long-live-moneyball/

Moneyball Assignment #1

1.0

Yun Mai
Gurpreet Singh
Dhananjay Kumar
Chirag Vithalani
CUNY (City University of New York)
New York, NY 10017
Advisor / Guide: Marcus Ellis

TABLE OF CONTENTS

DATA EXPLORATION

DATA PREPARATION

MODEL BUILDING

MODEL SELECTION

APPENDIX

Overview

1. DATA EXPLORATION

Describing the size and the variables in the moneyball training data set.

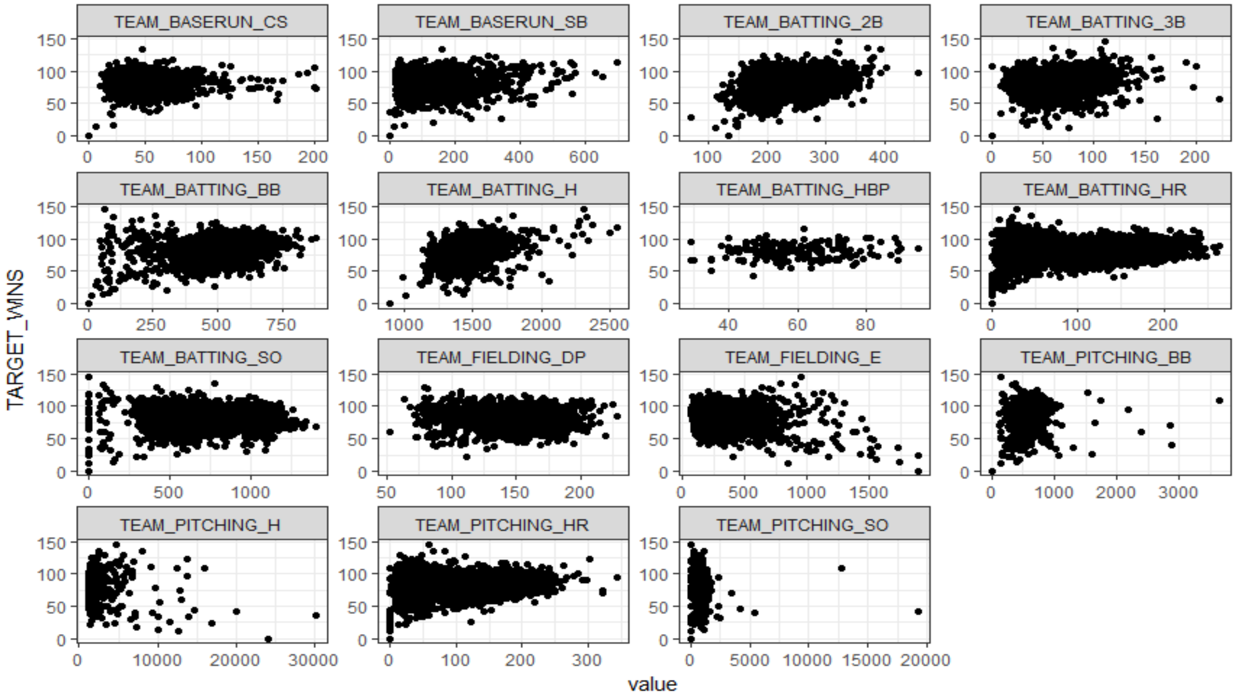
1.1. Data Description

The dataset consists of 17 variables and 2276 rows. Column INDEX is assigning a record number to each entry in the dataset and will be excluded from the analysis. The structure of data is integer, we will proceed with the analysis with the structure unchanged. The reference file describes the variables and their impact on the wins.

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of 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_BB	Walks by batters	Positive Impact on Wins
TEAM_BATTING_HBP	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

Variable TARGET_WINS is our dependent variable and remaining 15 variables are the independent variables that will be analyzed for prediction to use in our models. So, *since winning is everything*, we plot all variables against winning.



This graph shows how each variable impact to dependent variable TARGET_WINS

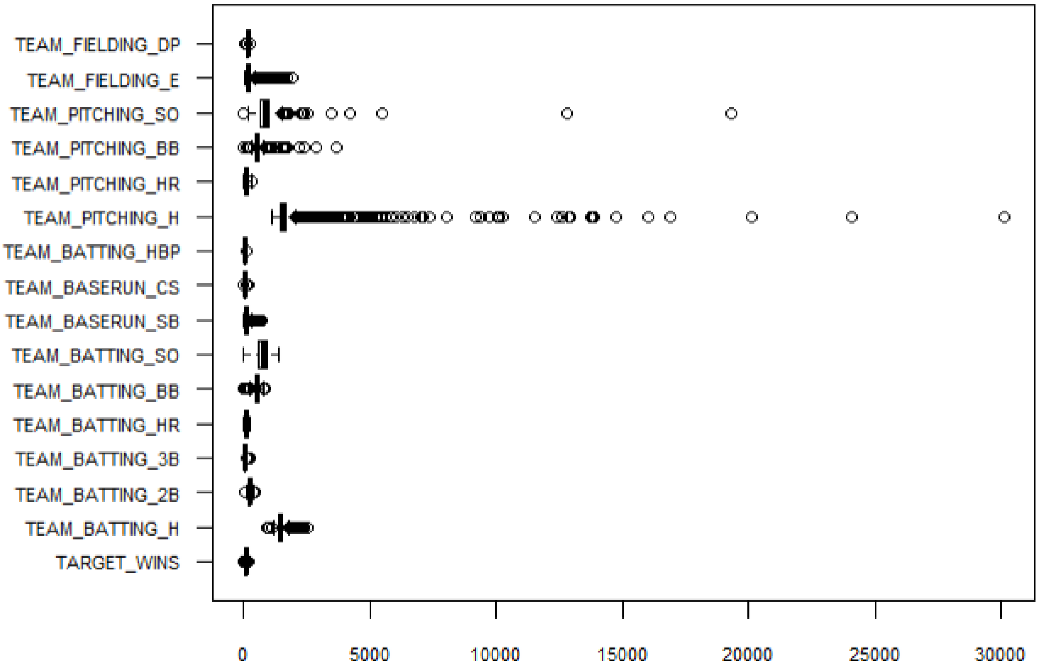
1.2 Data Summary

	mean	sd	med	min	max	range	skew	kurtosis	se
TARGET_WINS	80.79	15.75	82	0	146	146	-0.4	1.03	0.33
TEAM_BATTING_H	1469.27	144.59	1454	891	2554	1663	1.57	7.28	3.03
TEAM_BATTING_2B	241.25	46.8	238	69	458	389	0.22	0.01	0.98
TEAM_BATTING_3B	55.25	27.94	47	0	223	223	1.11	1.5	0.59
TEAM_BATTING_HR	99.61	60.55	102	0	264	264	0.19	-0.96	1.27
TEAM_BATTING_BB	501.56	122.67	512	0	878	878	-1.03	2.18	2.57
TEAM_BATTING_SO	735.61	248.53	750	0	1399	1399	-0.3	-0.32	5.33
TEAM_BASERUN_SB	124.76	87.79	101	0	697	697	1.97	5.49	1.9
TEAM_BASERUN_CS	52.8	22.96	49	0	201	201	1.98	7.62	0.59
TEAM_BATTING_HBP	59.36	12.97	58	29	95	66	0.32	-0.11	0.94
TEAM_PITCHING_H	1779.21	1406.84	1518	1137	30132	28995	10.33	141.84	29.49
TEAM_PITCHING_HR	105.7	61.3	107	0	343	343	0.29	-0.6	1.28
TEAM_PITCHING_BB	553.01	166.36	536.5	0	3645	3645	6.74	96.97	3.49
TEAM_PITCHING_SO	817.73	553.09	813.5	0	19278	19278	22.17	671.19	11.86
TEAM_FIELDING_E	246.48	227.77	159	65	1898	1833	2.99	10.97	4.77

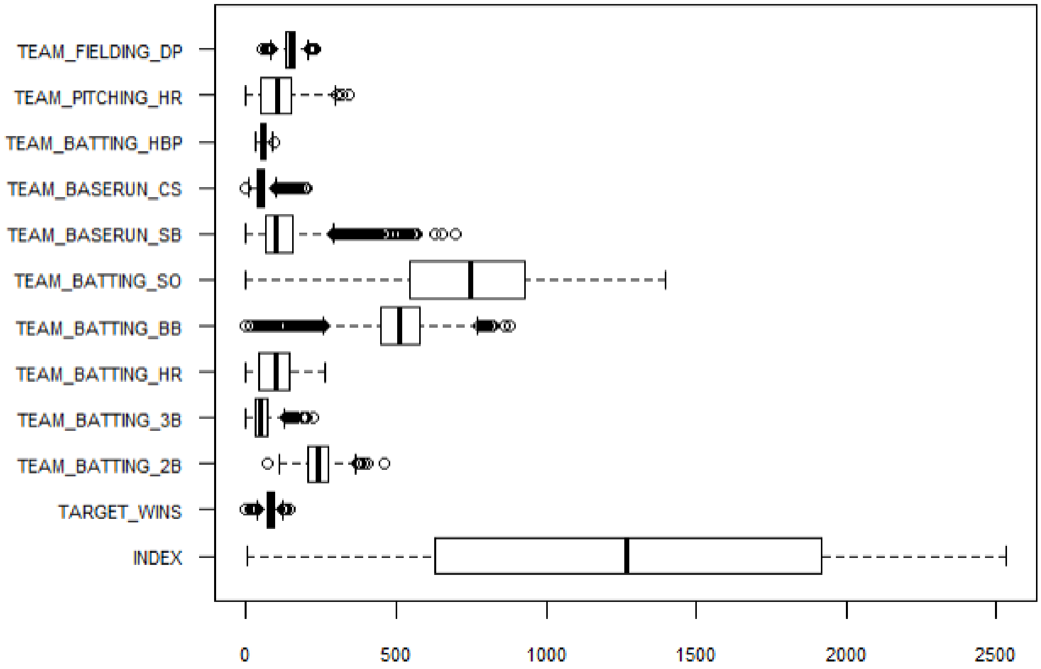
TEAM_FIELDING_DP	146.39	26.23	149	52	228	176	-0.39	0.18	0.59

1.3. Data Outliers

From the boxplot for all the variables, we see clearly that variables TEAM_PITCHING_H, TEAM_PITCHING_BB, TEAM_PITCHING_SO and TEAM_FIELDING_E contain the outliers.



Removing above variables with obvious outliers, we will be able to see other variables better. As below boxplot shown, there are no outliers in the rest variables. We will handle the outliers in the data preparation section.

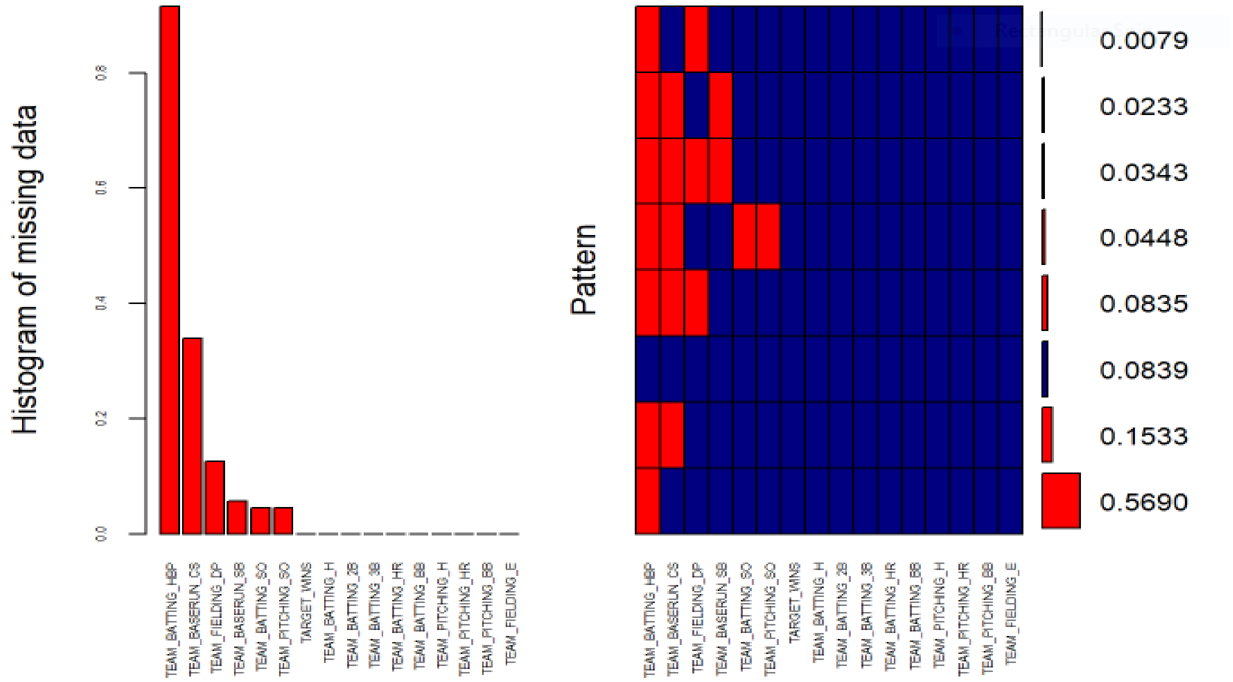


1.4. Missing Values (NA's)

We then check what percentage of data are missing in features and records. Defining the summary of data, we included the number of NA's in all the variables. The variables with missing values are listed in the following table. Usually a safe maximum threshold for missing values is 5% of the total for large datasets. So we will check whether the missing data for either predictor variables or samples/records is more than 5%.

predictors	missing_values	missing_values(%)
TEAM_BATTING_SO	102	4.5
TEAM_BASERUN_SB	131	5.8
TEAM_BASERUN_CS	772	33.9
TEAM_BATTING_HBP	2085	91.6
TEAM_PITCHING_SO	102	4.5
TEAM_FIELDING_DP	286	12.6

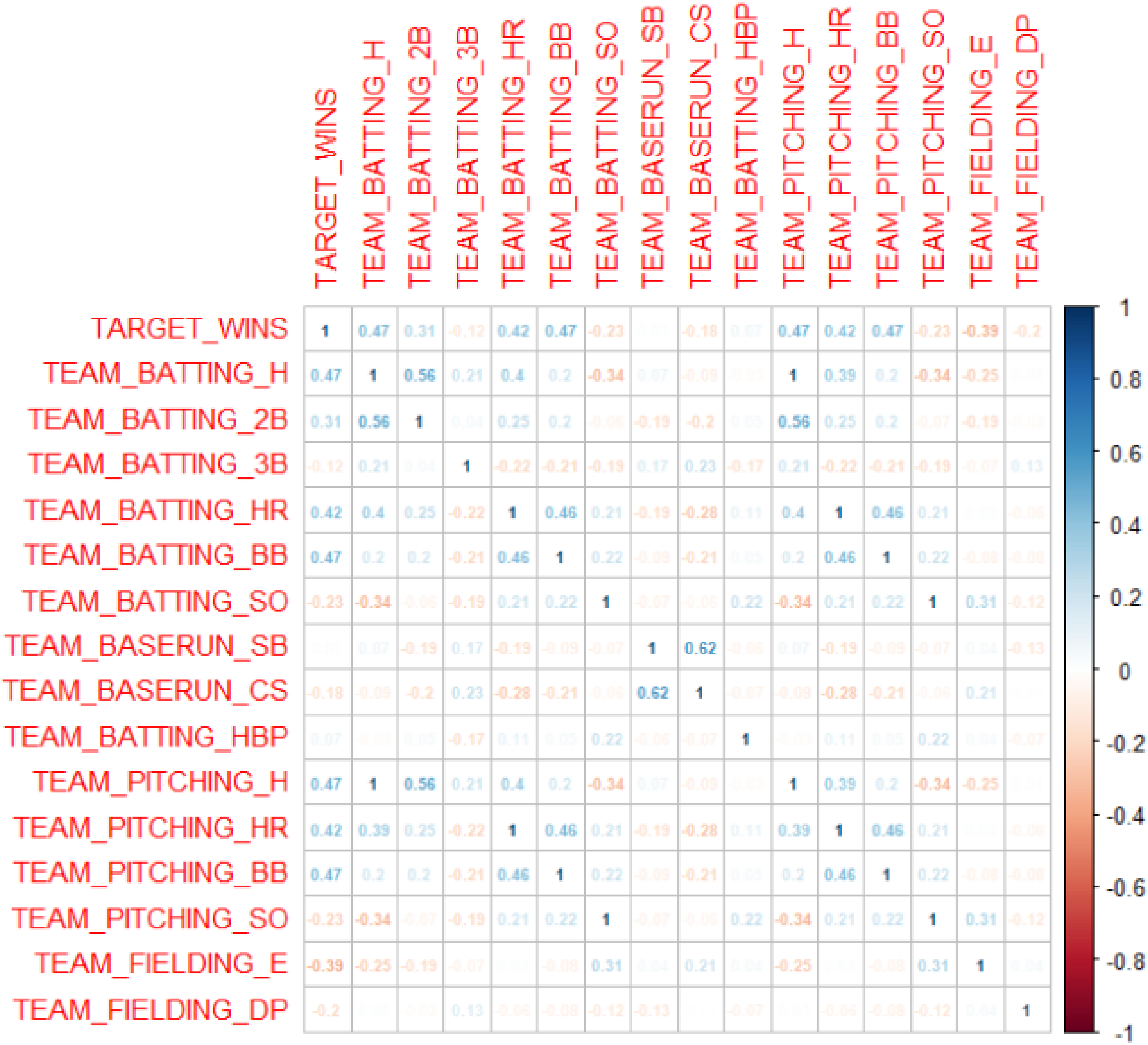
We see that variables TEAM_BASERUN_SB, TEAM_BASERUN_CS, TEAM_BATTING_HBP and TEAM_FIELDING_DP missed more than 5% data points. Since we can not get more measurements when doing the regression analysis, we may consider drop these features. For now we will only drop TEAM_BATTING_HBP which missed 92% data points.). Including the variable in the analysis might not be the best approach. In addition, imputation of the variables with large percentage of NA's might not be an effective way to handle NA's. We will keep other variables for now and decide whether to drop the three other variables later after checking whether they have significance in linear regression model. We may treat these variables with missing values by imputation.



The plot show that almost 8% of the samples are not missing any information, 56% samples are missing the TEAM_BATTING_HBP value. TEAM_BASERUN_CS, TEAM_FIELDING_DP, TEAM_BASERUN_SB, TEAM_BATTING_SO, and TEAM_PITCHING_SO show different missing patterns.

1.5. Collinearity

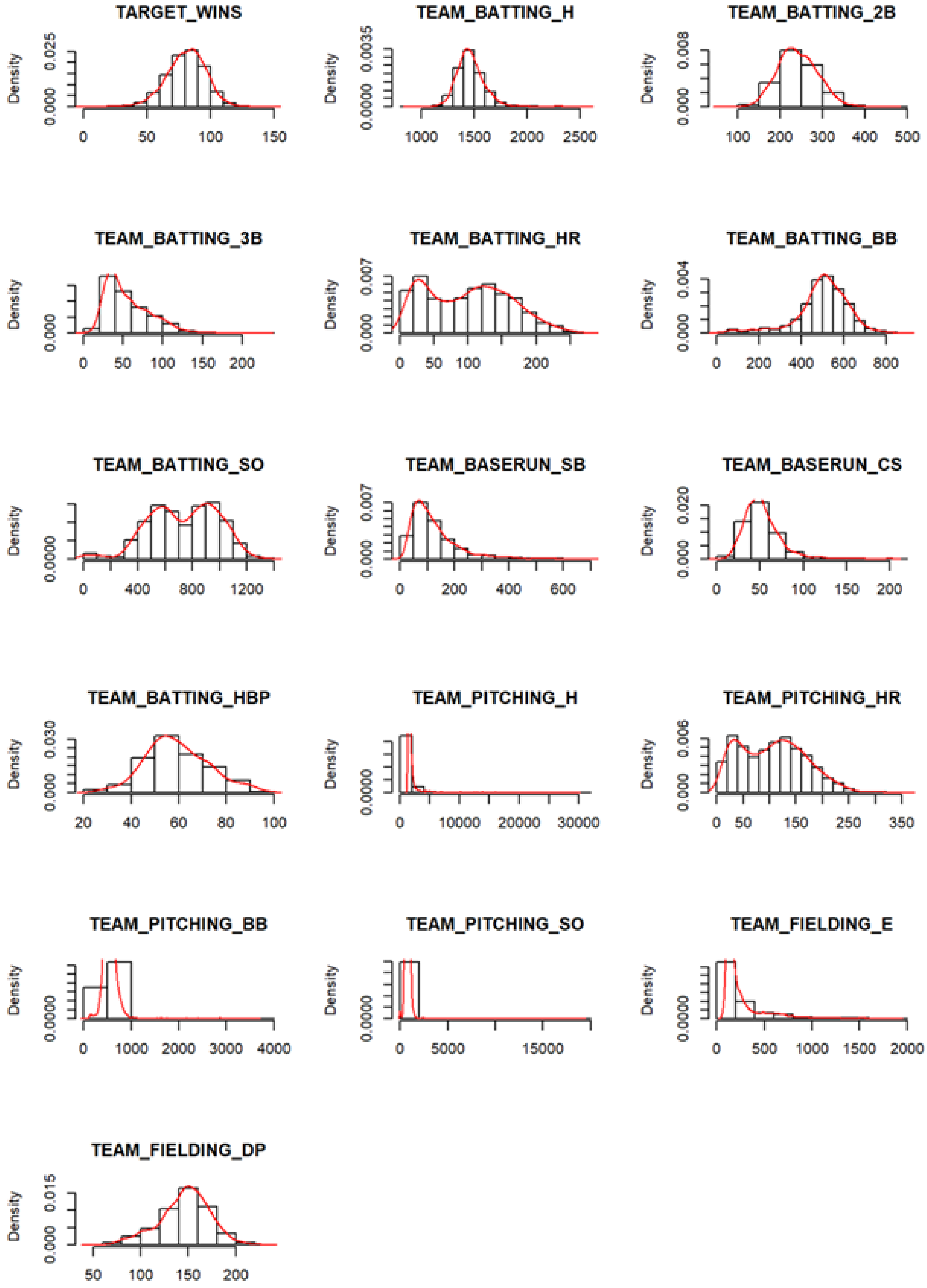
Based on the correlation plot, we can see that base hits, home runs, walks and strikeouts from the batting team and those from the pitching teams formed strong correlated (100%) pairs. So we can omit the redundant variables but only keep those data from the batting team. Data Preparation section leads us in the removal of four variables from analysis: TEAM_PITCHING_H, TEAM_PITCHING_HR, TEAM_PITCHING_BB and TEAM_PITCHING_SO. We will move forward with remaining variables for building the models.



	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
TARGET_WINS	1	0.47	0.31	-0.12	0.42	0.47	-0.23	0.01	-0.18	0.07	0.47	0.42	0.47	-0.23	-0.39	-0.2
TEAM_BATTING_H	0.47	1	0.56	0.21	0.4	0.2	-0.34	0.07	-0.09	-0.03	1	0.39	0.2	-0.34	-0.25	0.02
TEAM_BATTING_2B	0.31	0.56	1	0.04	0.25	0.2	-0.06	-0.19	-0.2	0.05	0.56	0.25	0.2	-0.07	-0.19	-0.02
TEAM_BATTING_3B	-0.12	0.21	0.04	1	-0.22	-0.21	-0.19	0.17	0.23	-0.17	0.21	-0.22	-0.21	-0.19	-0.07	0.13
TEAM_BATTING_HR	0.42	0.4	0.25	-0.22	1	0.46	0.21	-0.19	-0.28	0.11	0.4	1	0.46	0.21	0.02	-0.06
TEAM_BATTING_BB	0.47	0.2	0.2	-0.21	0.46	1	0.22	-0.09	-0.21	0.05	0.2	0.46	1	0.22	-0.08	-0.08
TEAM_BATTING_SO	-0.23	-0.34	-0.06	-0.19	0.21	0.22	1	-0.07	-0.06	0.22	-0.34	0.21	0.22	1	0.31	-0.12
TEAM_BASERUN_SB	0.01	0.07	-0.19	0.17	-0.19	-0.09	-0.07	1	0.62	-0.06	0.07	-0.19	-0.09	-0.07	0.04	-0.13
TEAM_BASERUN_CS	-0.18	-0.09	-0.2	0.23	-0.28	-0.21	-0.06	0.62	1	-0.07	-0.09	-0.28	-0.21	-0.06	0.21	[illegible]
TEAM_BATTING_HBP	0.07	-0.03	0.05	-0.17	0.11	0.05	0.22	-0.06	-0.07	1	-0.03	0.11	0.05	0.22	0.04	-0.07
TEAM_PITCHING_H	0.47	1	0.56	0.21	0.4	0.2	-0.34	0.07	-0.09	-0.03	1	0.39	0.2	-0.34	-0.25	0.02
TEAM_PITCHING_HR	0.42	0.39	0.25	-0.22	1	0.46	0.21	-0.19	-0.28	0.11	0.39	1	0.46	0.21	0.02	-0.06
TEAM_PITCHING_BB	0.47	0.2	0.2	-0.21	0.46	1	0.22	-0.09	-0.21	0.05	0.2	0.46	1	0.22	-0.08	-0.08
TEAM_PITCHING_SO	-0.23	-0.34	-0.07	-0.19	0.21	0.22	1	-0.07	-0.06	0.22	-0.34	0.21	0.22	1	0.31	-0.12
TEAM_FIELDING_E	-0.39	-0.25	-0.19	-0.07	0.02	-0.08	0.31	0.04	0.21	0.04	-0.25	0.02	-0.08	0.31	1	0.04
TEAM_FIELDING_DP	-0.2	0.02	-0.02	0.13	-0.06	-0.08	-0.12	-0.13	[illegible]	-0.07	0.02	-0.06	-0.08	-0.12	0.04	1

1.6. Normality

Normality check was performed on independent variables. We used the density plots and histograms for independent variables to check for normality assumption.


TARGET_WINS
Density
0.000
0.025
0
50
100
150
TEAM_BATTING_H
Density
0.0000
0.0035
1000
1500
2000
2500
TEAM_BATTING_2B
Density
0.000
0.008
100
200
300
400
500
TEAM_BATTING_3B
Density
0.000
0
50
100
150
200
TEAM_BATTING_HR
Density
0.000
0.007
0
50
100
200
TEAM_BATTING_BB
Density
0.000
0.004
0
200
400
600
800
TEAM_BATTING_SO
Density
0.0000
0
400
800
1200
TEAM_BASERUN_SB
Density
0.000
0.007
0
200
400
600
TEAM_BASERUN_CS
Density
0.000
0.020
0
50
100
150
200
TEAM_BATTING_HBP
Density
0.000
0.030
20
40
60
80
100
TEAM_PITCHING_H
Density
0.0000
0
10000
20000
30000
TEAM_PITCHING_HR
Density
0.000
0.006
0
50
150
250
350
TEAM_PITCHING_BB
Density
0.0000
0
1000
2000
3000
4000
TEAM_PITCHING_SO
Density
0.0000
0
5000
15000
TEAM_FIELDING_E
Density
0.0000
0
500
1000
1500
2000
TEAM_FIELDING_DP
Density
0.000
0.015
50
100
150
200

Based on the plots, we assumed variables TEAM_BATTING_3B,TEAM_BATTING_HR, TEAM_BASERUN_SB, TEAM_PITCHING_H, TEAM-PITCHING_HR, TEAM_PITCHING_BB, TEAM_PITCHING_SO and PITCHING_E can not be assumed to follow normality. Since TEAM_PITCHING_H, TEAM-PITCHING_HR, TEAM_PITCHING_BB, TEAM_PITCHING_SO have be omitted because of collinearity issue, the rest 4 variables needed to be transformed in order to build our model if they are useful modeling. The transformation of the variables will be performed in the data preparation section.

2. DATA PREPARATION

2.1 Remove predictor variables

According to the results of checking the missing data, examine the collinearity, some predictor variables will be removed.

- ***TEAM_BATTING_HBP:*** Too many missing data points(92%).
- ***TEAM_PITCHING_H,TEAM_PITCHING_HR,TEAM_PITCHING_BB and TEAM_PITCHING_SO*** will be removed from the data because of the collinearities.
- ***TEAM_BASERUN_CS:*** Cause stolen data is highly related (0.62) to stolen base (TEAM_BASERUN_SB) data. Moreover, this variable missed 33.9% data points. But this variable will be added back in one of the models later for generating new variable. The corresponding data set is called train_clean2_im
- ***TEAM_FIELDING_DP:*** We will keep this variable because it is important to building the model (p-value<0.05) even it missed 33% data points. We can impute the missing data before we build the model.
- ***TEAM_BASERUN_SB:*** We will keep this variable since the proportion of missing data is 5.8% which is barely above the 5% threshold. Also, one of the variables related to stole base TEAM_BASERUN_CS which has strong correlation to TEAM_BASERUN_SB will be removed. This variable could be useful for the modeling.

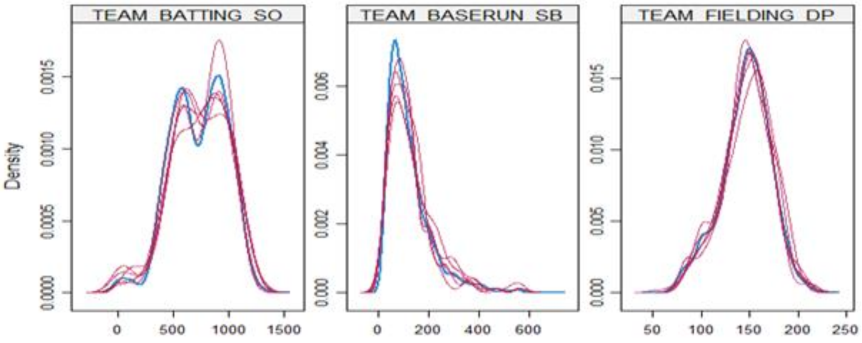
2.2 Remove records with too many missing data points

As far as the samples are concerned,there are 441(19%) sample misse 5% or more datapoints.So 5% would be too strict for this data set. There are 78 records has more than 10% missing data. To compare whether imputation will improve the model,

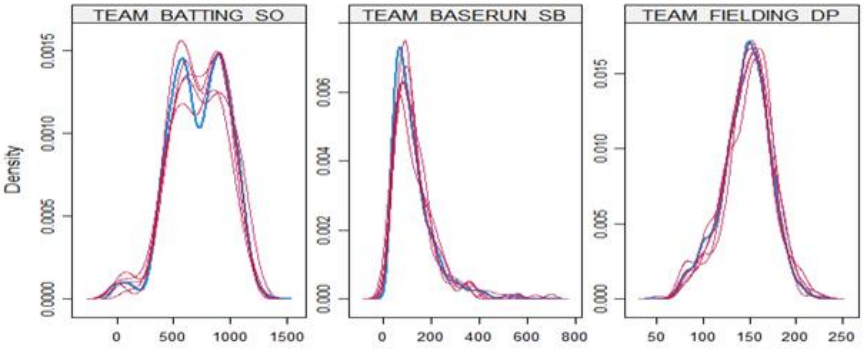
1. We will keep all samples and impute the missing data points and this data, called train_clean_im, set will serve for model #1.
2. We will remove the 78 records and build the model based the rest data and this data set, called train_dlrd_im, will serve for model #2.

2.3 Impute the missing values

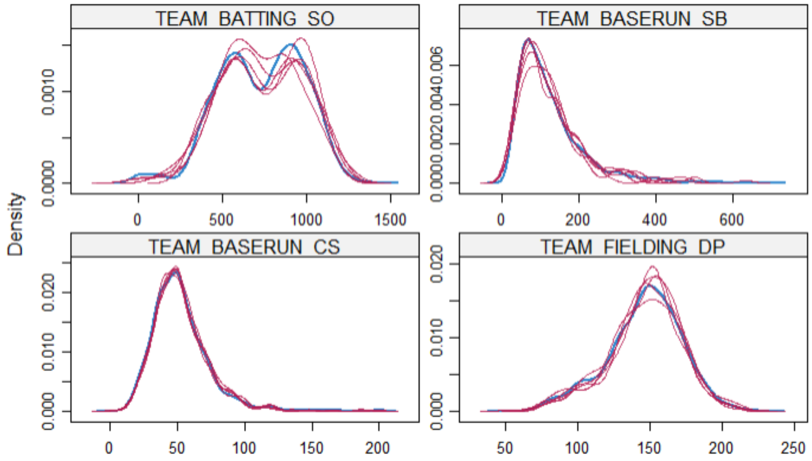
After removing the variables that miss more than 5% data points, there are still some variables missed less than 5% data points. We then impute the missing data for further analysis.



Density plots for imputed values and observed values in train_clean_im.



Density plots for imputed values and observed values in train_dlrd_im.



Density plots for imputed values and observed values in train_clean2_im.

Sample method has been used in the imputation. As shown in the above density plots, the distribution of the five imputed data sets which are presented as blue is similar to the density of the observed data presented as blue.

To the data set **train_clean_im** for for model #1 #3 and #4, 102 missing values of "TEAM_BATTING_SO" were imputed. 131 missing values of "TEAM_BASERUN_SB" were imputed. 286 missing values of "TEAM_FIELDING_DP" were imputed.

To the data set **train_dlrd_im** used for model #2, 100 missing values of "TEAM_BATTING_SO" were imputed. 121 missing values of "TEAM_BASERUN_SB" were imputed. 269 missing values of "TEAM_FIELDING_DP" were imputed.

For the data set **train_clean2_im** used for model #5, 102 missing values of "TEAM_BATTING_SO" were imputed. 131 missing values of "TEAM_BASERUN_SB" were imputed. 286 missing values of "TEAM_FIELDING_DP" were imputed, and 772 missing values of "TEAM_BASERUN_CS" were imputed.

2.4 Transform the data

2.4.1 Create a new derived variable

In the baseball, a run is scored either when the player hits a home run or when he advances around first, second and third and return to the home plate safely. Some combination of all successful singles, doubles, triples and walks will contribute to the chance of scoring. The baseball writer and statistician, Bill James, invented Runs created (RC) to estimate the number of runs a batter contributes to his team. We will use the base formula and two advanced formula to generate two new variables.

Basic runs created

$$RC = \frac{(H+BB)\times TB}{AB+BB}$$

"Stolen base" version of runs created:

$$RC = \frac{(H+BB-CS)(TB+0.55\times SB)}{AB+BB}$$

H: hits, also called a base hits

TB: total bases (TBs) is the number of bases a player has gained with hits.

$$totalbases(TB) = 1\times Singles + 2\times Doubles + 3\times Triples + 4\times Homeruns$$

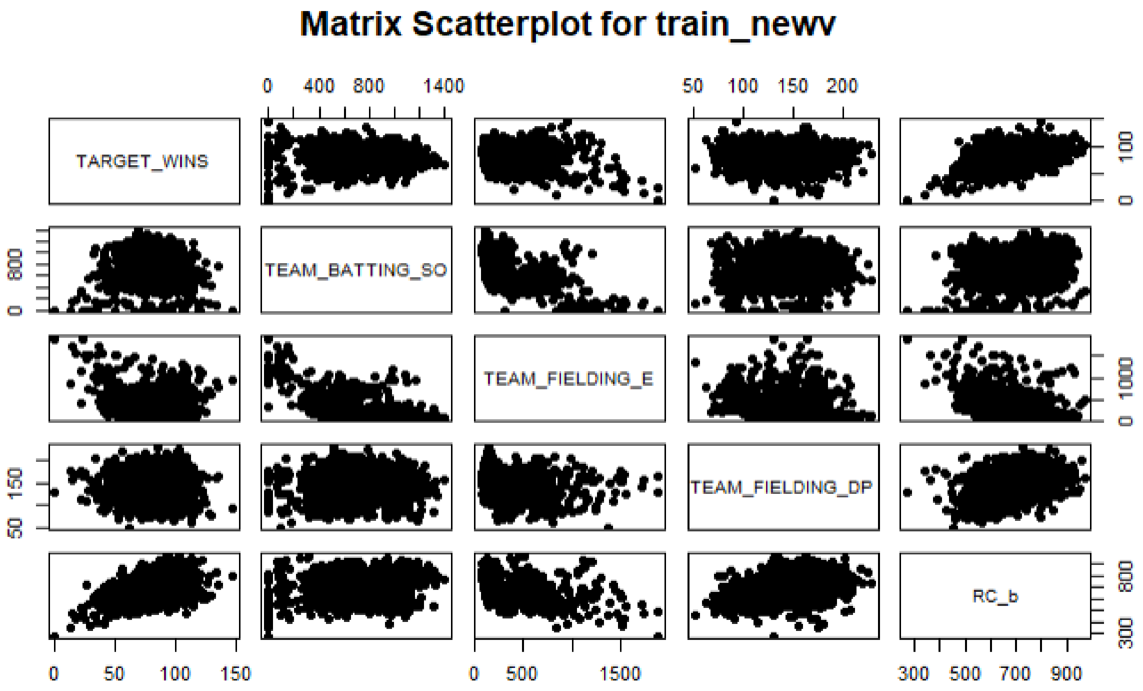
or

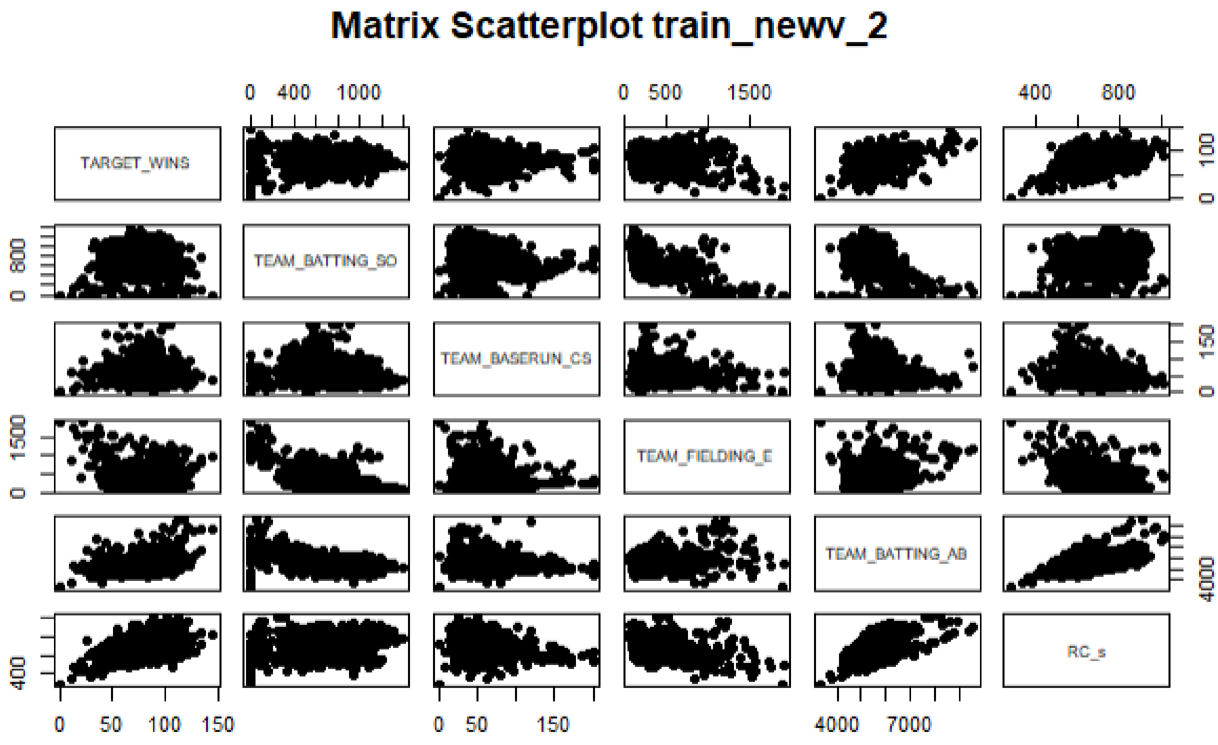
$$totalbases(TB) = H + Doubles + 2\times Triples + 3\times Homeruns$$

AB: an at bat (AB) or time at bat is a batter's turn batting against a pitcher. A batter is credited with an at bat only if that plate appearance does not have base on balls, hit by a pitch, or the hitter hits a sacrifice fly or a sacrifice bunt, etc. We do not have at bat data but we can estimate the value from the batting average (BA). batting average is defined by the number of hits divided by at bats. From the Wikipedia we know that the all-time league batting average in Major League Baseball is between .260 and .275. So we can estimate AB as H/((.26+.275)/2).

The new variables and the components of the formula are clearly collinear variables. So we remove those variables consist the right side of the formula but keep the new variable runs created and the left variables as predictor. variance inflation factors (VIF) will be used to examine whether there are no collinearity issues.

VIF test shows there is no collinearities between predictors. The new data set train_newv which contains 4 predictor variables, TEAM_BATTING_SO, TEAM_FIELDING_E, TEAM_FIELDING_DP and RC_b (basic runs created) will be used for model #4. The new data set train_newv_2 which contains 5 predictor variables,TEAM_BATTING_SO, TEAM_BASERUN_CS, TEAM_FIELDING_E, TEAM_BATTING_AB and RC_s ("Stolen base" version of runs created) will be used for model #5. We then check the relationship among all variables in one single matrix plot for each of these two new data sets.

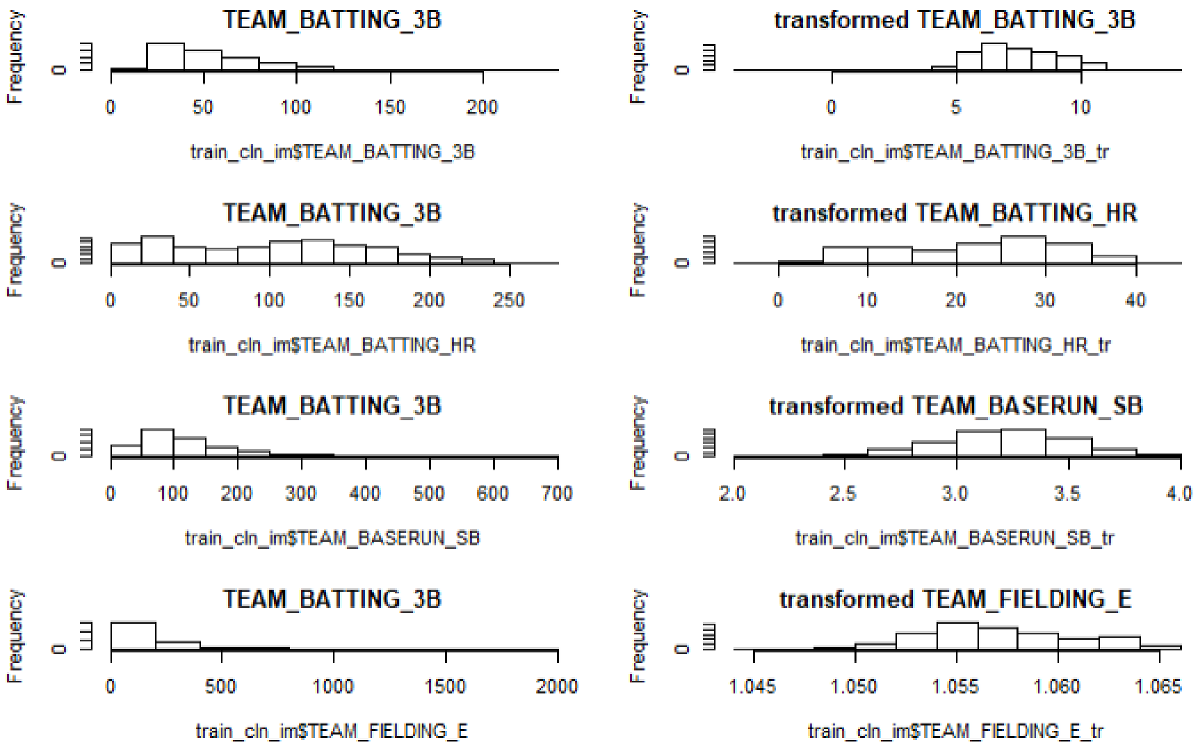


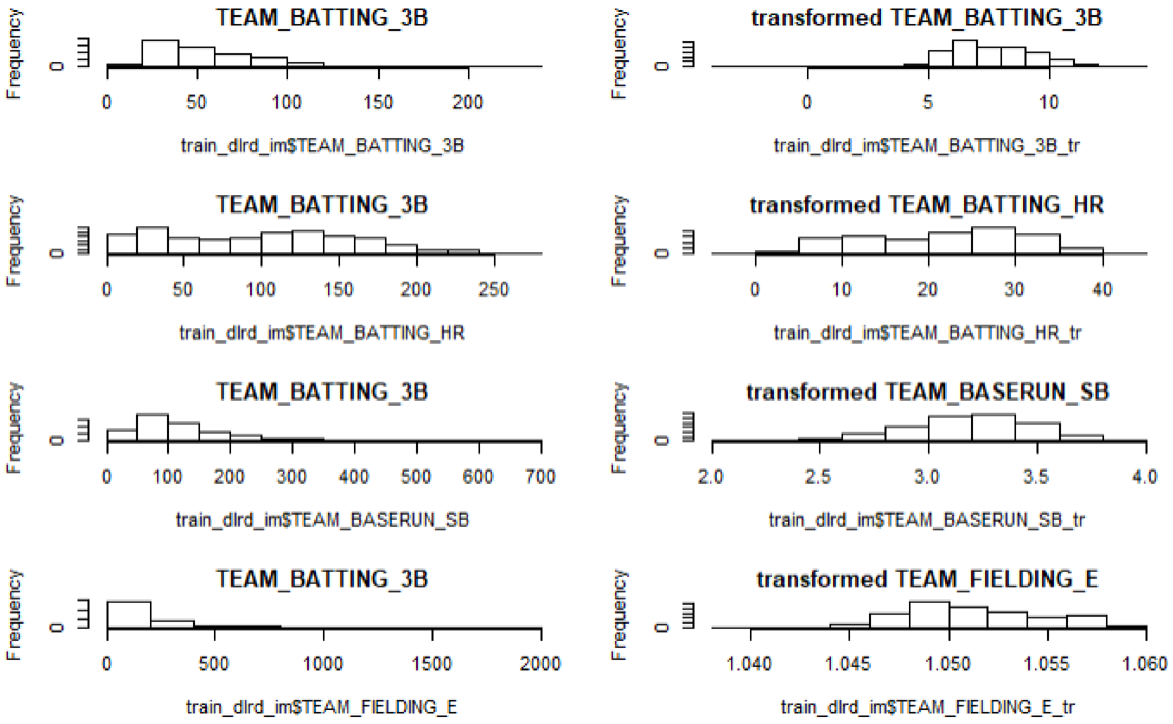


2.4.2 Box-Cox Transformation of predictors

After imputation and generating new variables, we will make transformation for variables have non-normality of distribution.

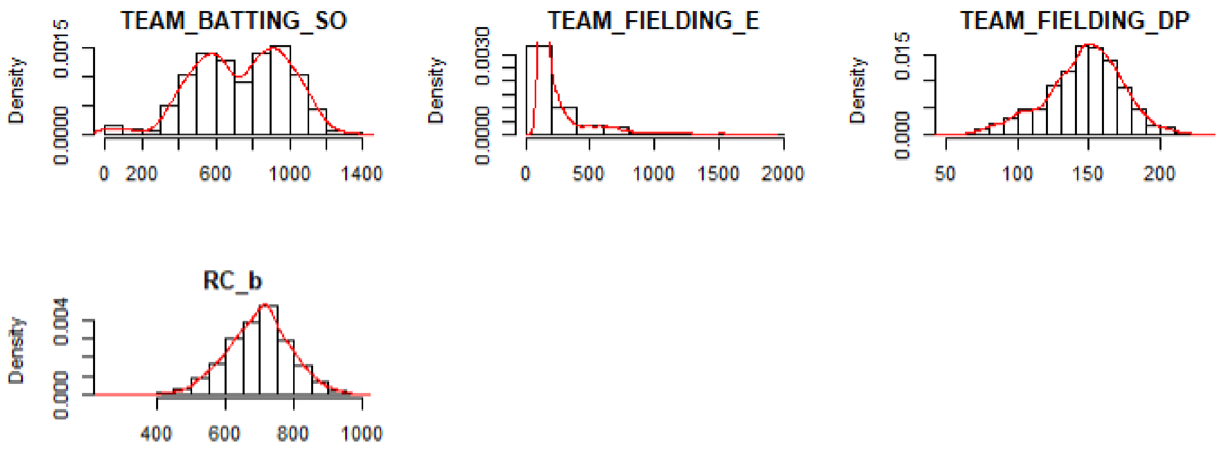
Base one the conclusion from the data exploration, the variables in train_cln_im and train_dlrd_im(deleting some records with >10% missing values) need to be transformed for basic linear regression model are: TEAM_BATTING_3B,TEAM_BATTING_HR, TEAM_BASERUN_SB and TEAM_FIELDING_E .


TEAM_BATTING_3B
Frequency
0
0
50
100
150
200
train_cln_im$TEAM_BATTING_3B
transformed TEAM_BATTING_3B
Frequency
0
0
5
10
train_cln_im$TEAM_BATTING_3B_tr
TEAM_BATTING_3B
Frequency
0
0
50
100
150
200
250
train_cln_im$TEAM_BATTING_HR
transformed TEAM_BATTING_HR
Frequency
0
0
10
20
30
40
train_cln_im$TEAM_BATTING_HR_tr
TEAM_BATTING_3B
Frequency
0
0
100
200
300
400
500
600
700
train_cln_im$TEAM_BASERUN_SB
transformed TEAM_BASERUN_SB
Frequency
0
2.0
2.5
3.0
3.5
4.0
train_cln_im$TEAM_BASERUN_SB_tr
TEAM_BATTING_3B
Frequency
0
0
500
1000
1500
2000
train_cln_im$TEAM_FIELDING_E
transformed TEAM_FIELDING_E
Frequency
0
1.045
1.050
1.055
1.060
1.065
train_cln_im$TEAM_FIELDING_E_tr

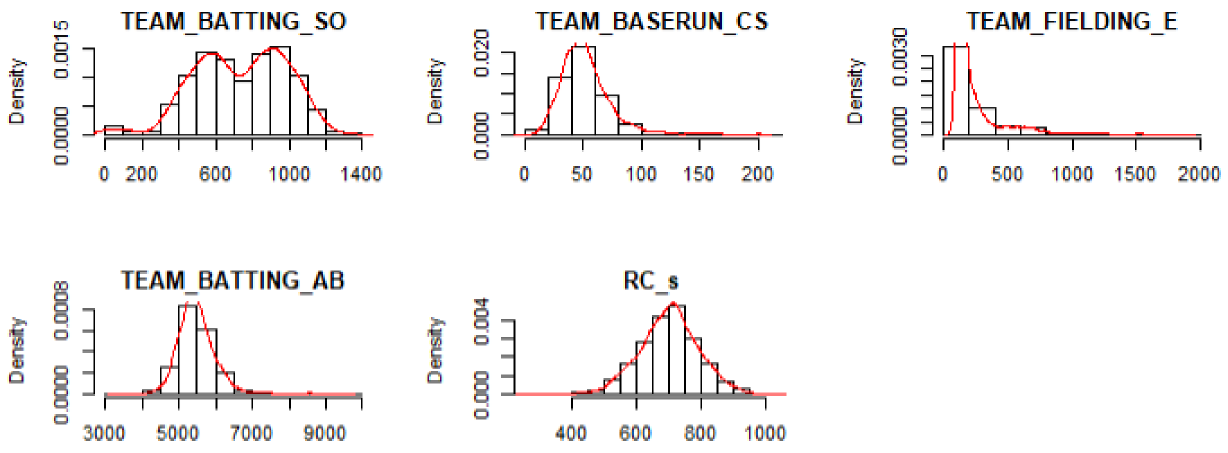

TEAM_BATTING_3B
Frequency
0
0
50
100
150
200
train_dlrd_im$TEAM_BATTING_3B
transformed TEAM_BATTING_3B
Frequency
0
0
5
10
train_dlrd_im$TEAM_BATTING_3B_tr
TEAM_BATTING_3B
Frequency
0
0
50
100
150
200
250
train_dlrd_im$TEAM_BATTING_HR
transformed TEAM_BATTING_HR
Frequency
0
0
10
20
30
40
train_dlrd_im$TEAM_BATTING_HR_tr
TEAM_BATTING_3B
Frequency
0
0
100
200
300
400
500
600
700
train_dlrd_im$TEAM_BASERUN_SB
transformed TEAM_BASERUN_SB
Frequency
0
2.0
2.5
3.0
3.5
4.0
train_dlrd_im$TEAM_BASERUN_SB_tr
TEAM_BATTING_3B
Frequency
0
0
500
1000
1500
2000
train_dlrd_im$TEAM_FIELDING_E
transformed TEAM_FIELDING_E
Frequency
0
1.040
1.045
1.050
1.055
1.060
train_dlrd_im$TEAM_FIELDING_E_tr

The Box-Cox transformed predictors variables TEAM_BATTING_3B,TEAM_BATTING_HR, TEAM_BASERUN_SB and TEAM_FIELDING_E now follow normal distribution.

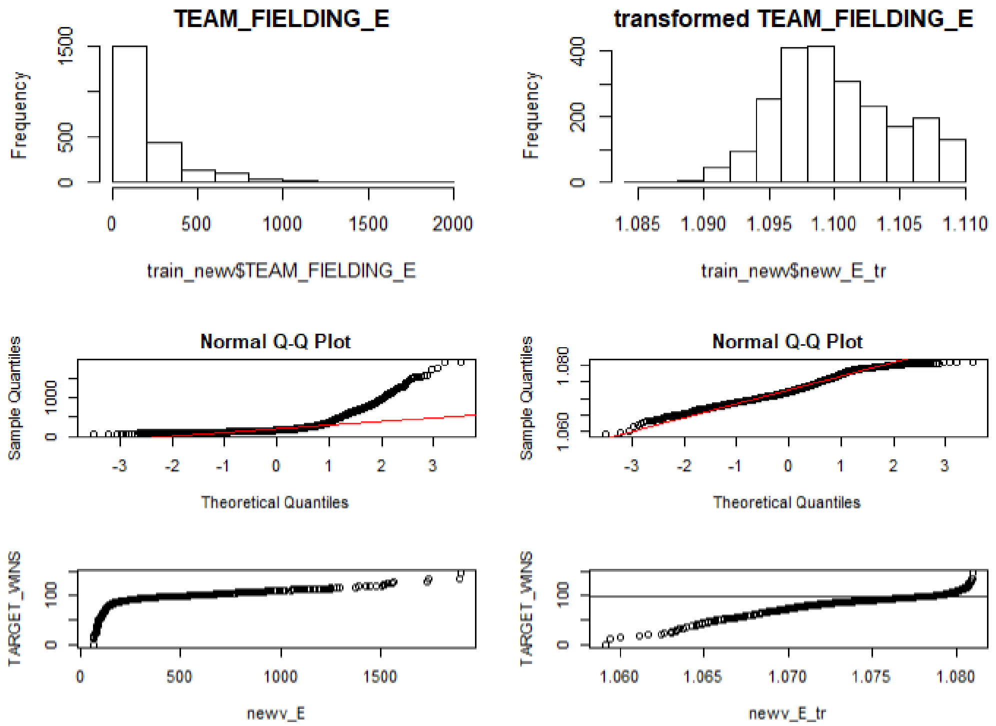
For basic and “stolen based” run created models, first check the distribution of the predictors in the corresponding data sets. From the figure below we can see that both basic run created values and “stolen based” run created values follow normal distribution. TEAM_FIELDING_E and TEAM_BASERUN_CS will be Box-Cox transformed.



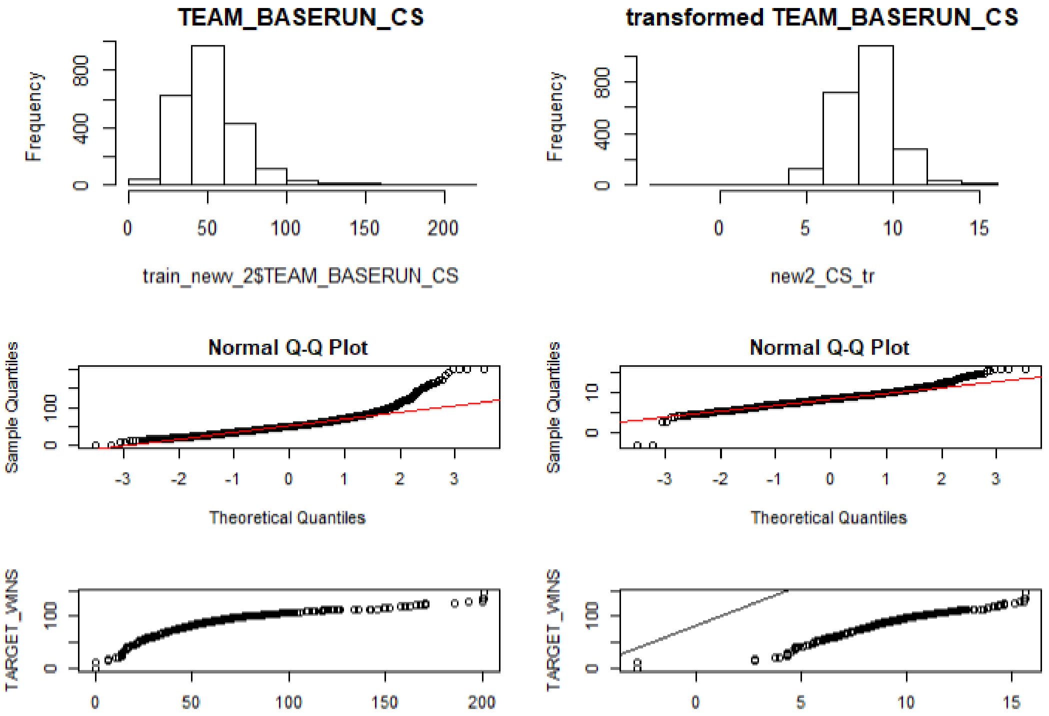
Distribution of predictor variables in train_newv data set.



Distribution of predictor variables in train_newv_2 data set.



Box-Cox transformed TEAM_FIELDING_E

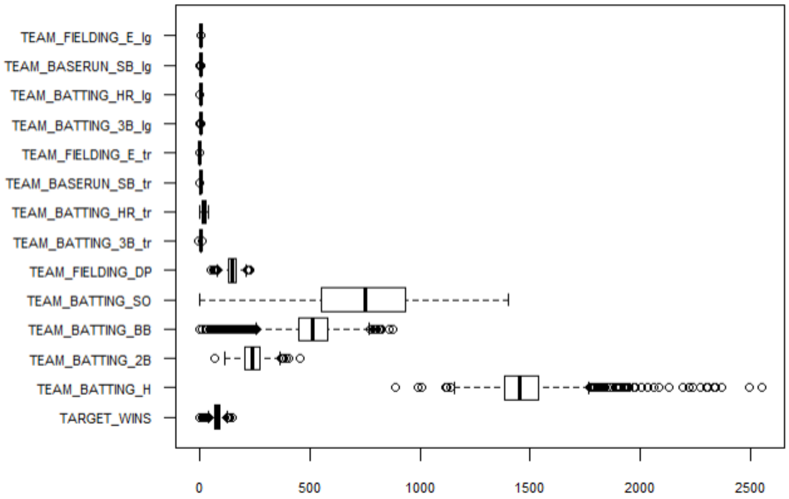


Box-Cox transformed TEAM_BASERUN_CS

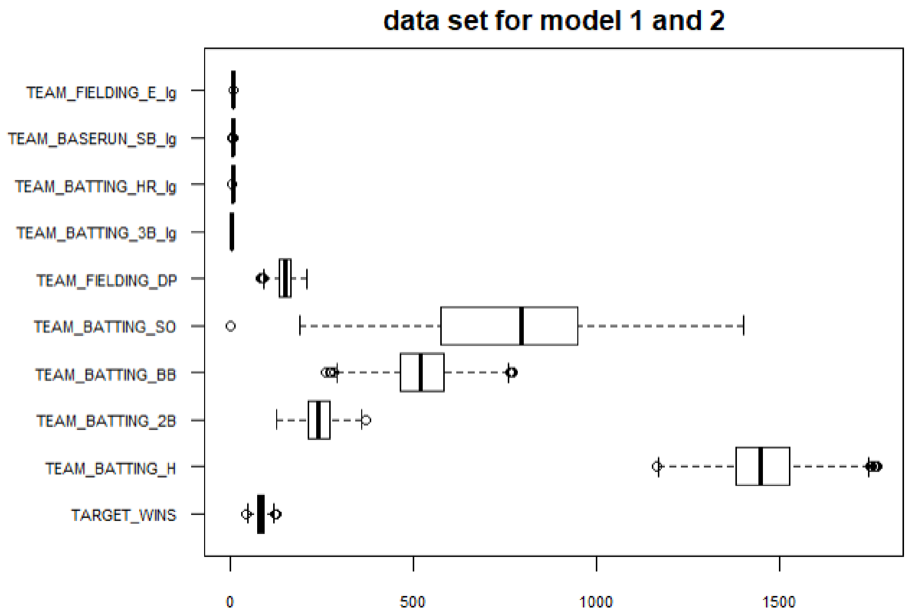
The Box-Cox transformed predictors variables TEAM_FIELDING_E and TEAM_BASERUN_CS follow normal distribution.

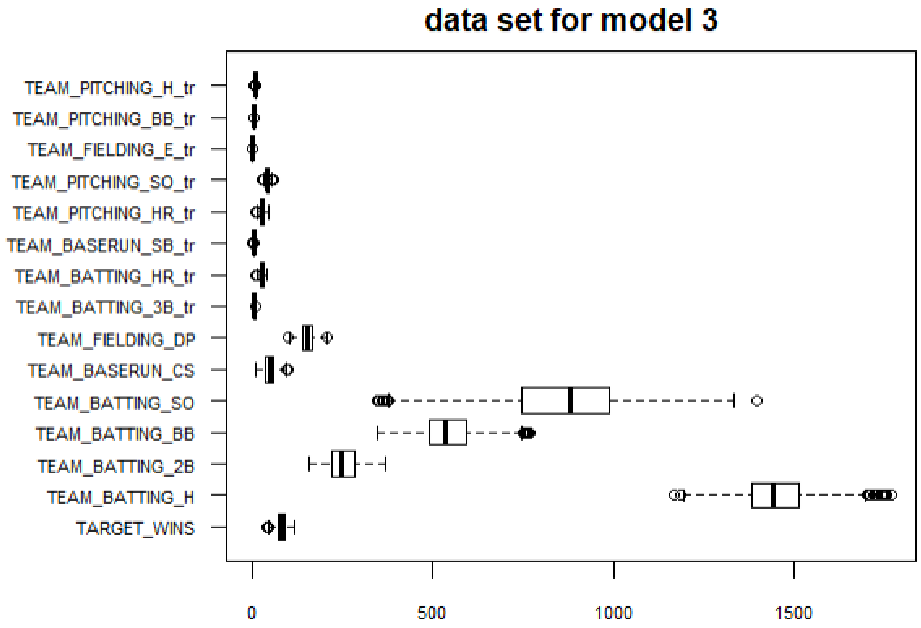
2.5 Handle Outliers in the dataset

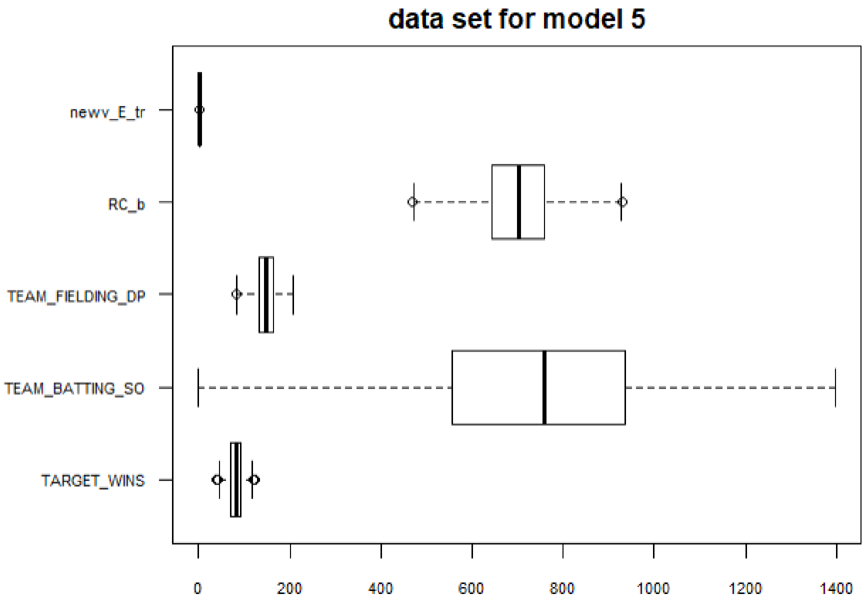
The boxplot show that data transformation process is dramatically improving the data.

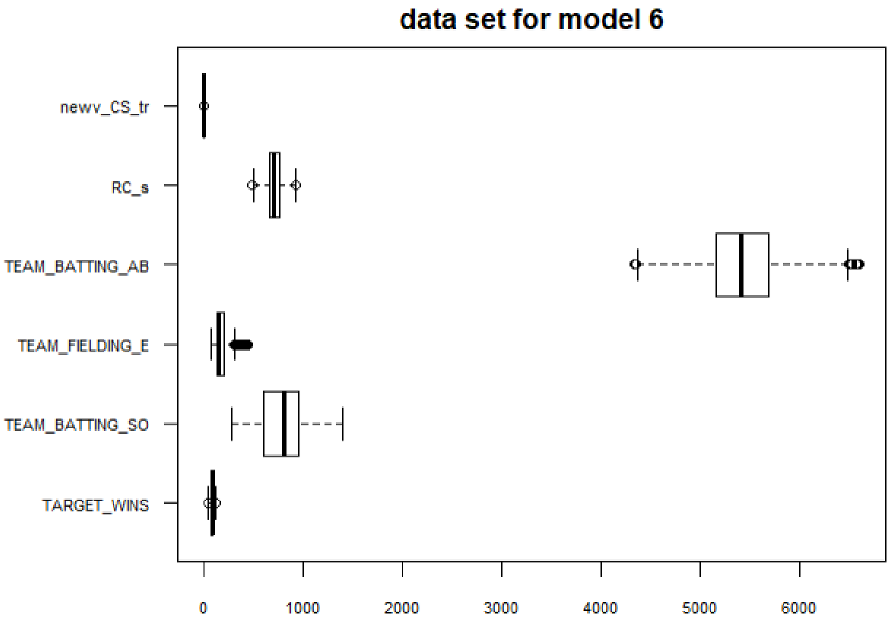


The values are beyond 1.5 x IQR are considered outliers . WE then remove outliers from the five datasets for modeling. The boxplot of the data sets below show that outliers were successfully removed.


data set for model 1 and 2
TEAM_FIELDING_E_lg
TEAM_BASERUN_SB_lg
TEAM_BATTING_HR_lg
TEAM_BATTING_3B_lg
TEAM_FIELDING_DP
TEAM_BATTING_SO
TEAM_BATTING_BB
TEAM_BATTING_2B
TEAM_BATTING_H
TARGET_WINS
0
500
1000
1500


data set for model 3
TEAM_PITCHING_H_tr
TEAM_PITCHING_BB_tr
TEAM_FIELDING_E_tr
TEAM_PITCHING_SO_tr
TEAM_PITCHING_HR_tr
TEAM_BASERUN_SB_tr
TEAM_BATTING_HR_tr
TEAM_BATTING_3B_tr
TEAM_FIELDING_DP
TEAM_BASERUN_CS
TEAM_BATTING_SO
TEAM_BATTING_BB
TEAM_BATTING_2B
TEAM_BATTING_H
TARGET_WINS
0
500
1000
1500



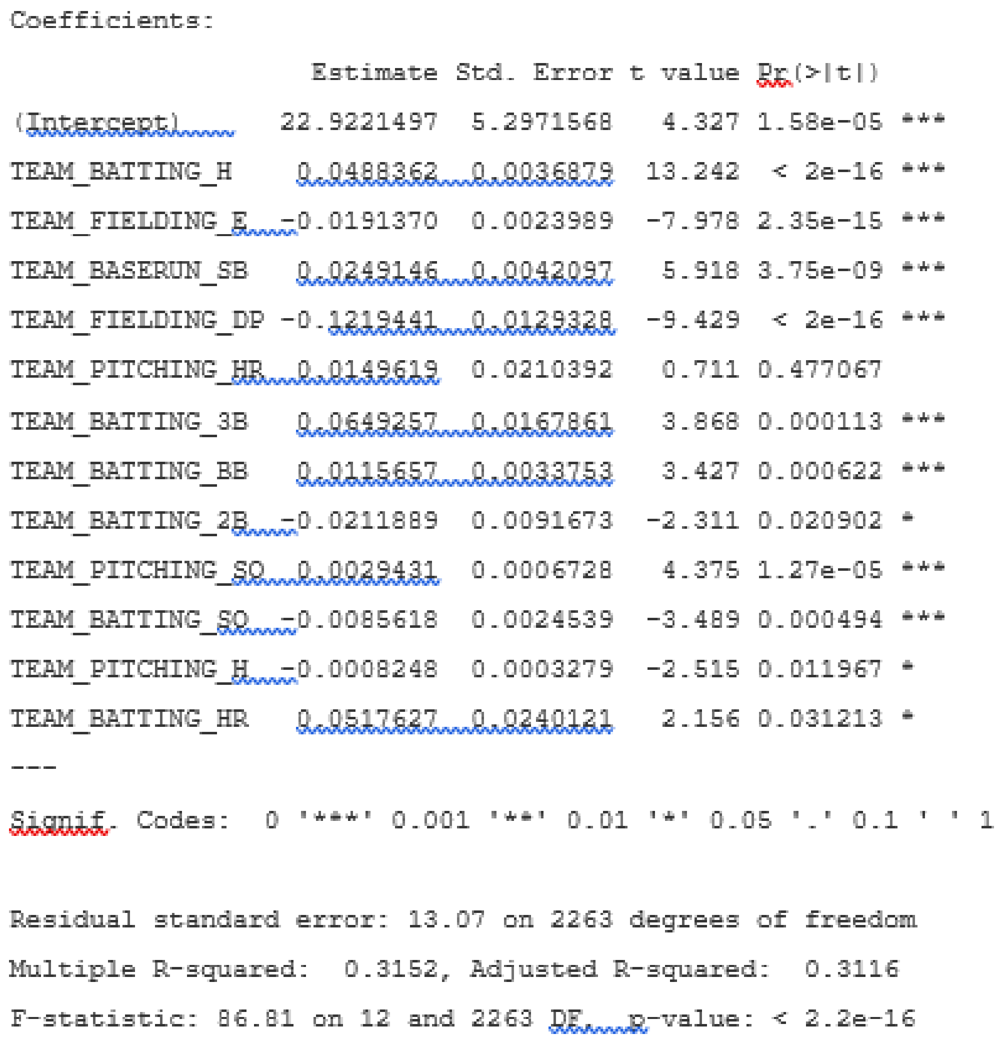


The data sets ready for used in the models are shown in the table below.

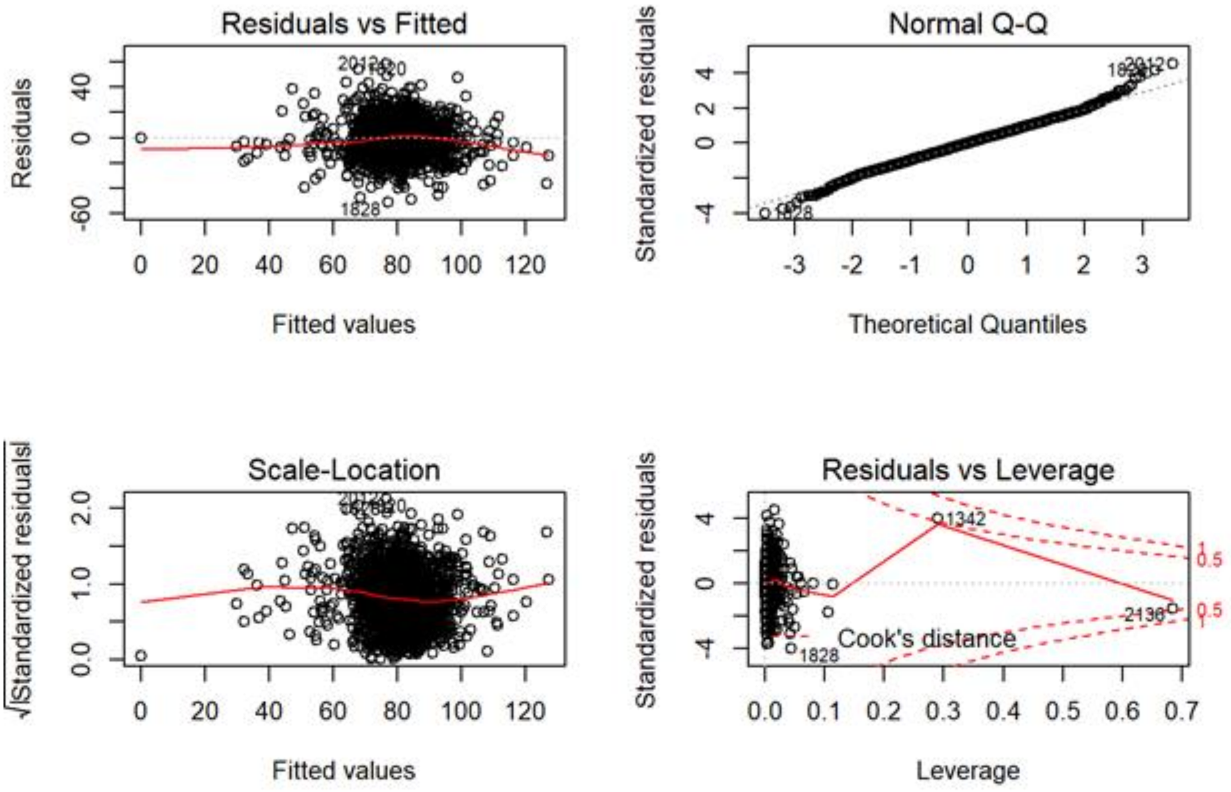
3. Build Models

Model-1. Log Transformation + Backward Elimination

Model-1 is created using the backward elimination process. The variables created using log transformation were used to build the model using Backward elimination. The summary is shown as below:



```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       22.9221497  5.2971568   4.327 1.58e-05 ***
TEAM_BATTING_H     0.0488362  0.0036879  13.242  < 2e-16 ***
TEAM_FIELDING_E   -0.0191370  0.0023989  -7.978 2.35e-15 ***
TEAM_BASERUN_SB    0.0249146  0.0042097   5.918 3.75e-09 ***
TEAM_FIELDING_DP -0.1219441  0.0129328  -9.429  < 2e-16 ***
TEAM_PITCHING_HR   0.0149619  0.0210392   0.711 0.477067
TEAM_BATTING_3B    0.0649257  0.0167861   3.868 0.000113 ***
TEAM_BATTING_BB    0.0115657  0.0033753   3.427 0.000622 ***
TEAM_BATTING_2B   -0.0211889  0.0091673  -2.311 0.020902 *
TEAM_PITCHING_SO   0.0029431  0.0006728   4.375 1.27e-05 ***
TEAM_BATTING_SO   -0.0085618  0.0024539  -3.489 0.000494 ***
TEAM_PITCHING_H   -0.0008248  0.0003279  -2.515 0.011967 *
TEAM_BATTING_HR    0.0517627  0.0240121   2.156 0.031213 *
---
Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.07 on 2263 degrees of freedom
Multiple R-squared:  0.3152, Adjusted R-squared:  0.3116
F-statistic: 86.81 on 12 and 2263 DF,  p-value: < 2.2e-16
```

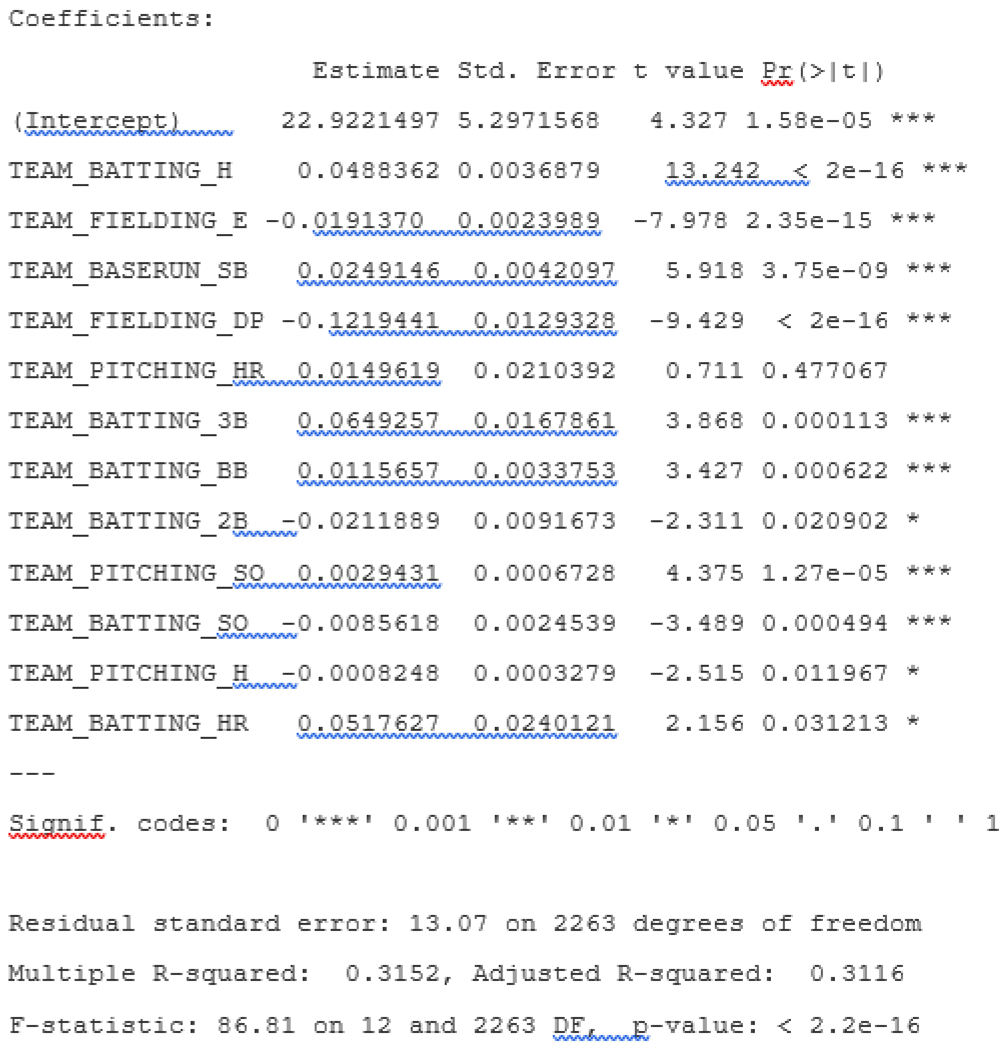


According to the theoretical effects of each predictors, the coefficients for all the predictors but TEAM_BATTING_2B make sense. TEAM_BATTING_2B coefficient surprises since it is believed that double base will improve a team's chances of scoring. All the assumptions are valid since the residuals have constant variance. There are no outliers beyond the Cook's distance that could have strong influence on the regression results.

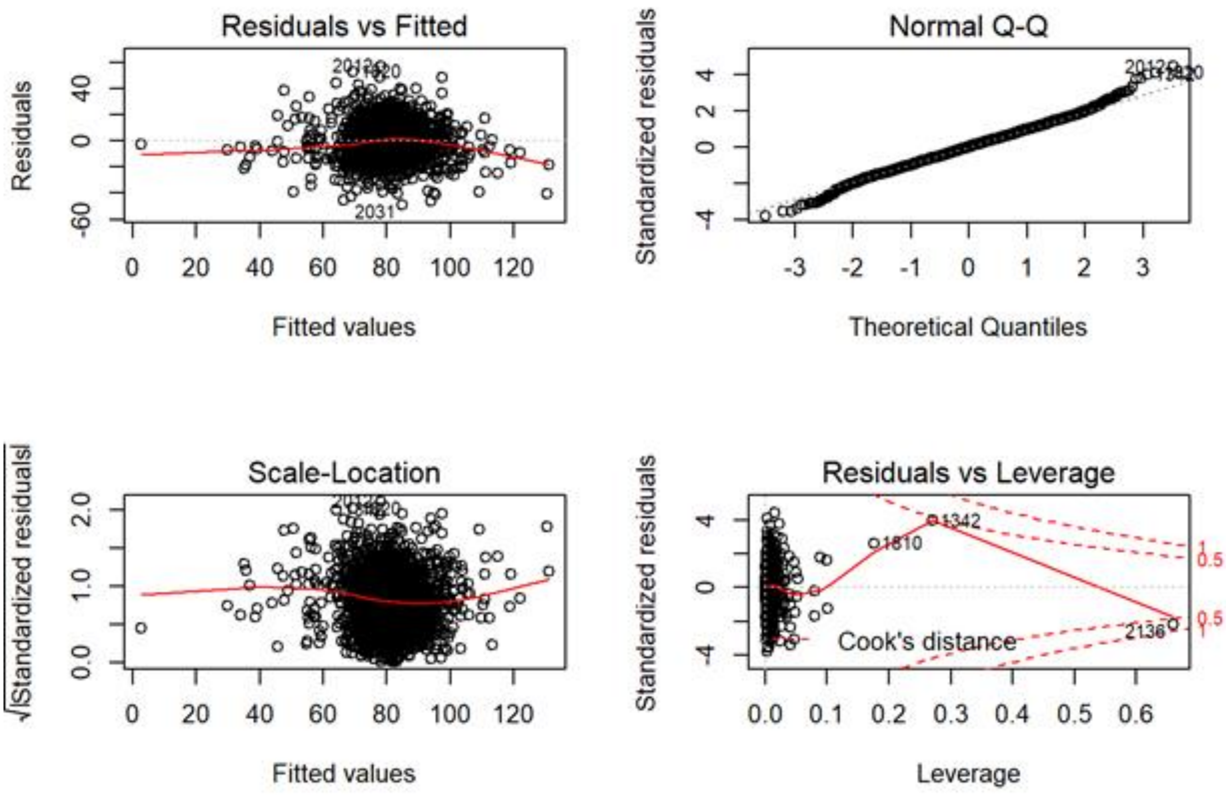
Model-2. Log Transformation + Forward Stepwise Selection

Model 2 was created using forward stepwise regression technique. The log transformed variables used in model 1 were picked to apply forward stepwise regression model.

A summary of model-2 is as follows:



```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      22.9221497 5.2971568   4.327 1.58e-05 ***
TEAM_BATTING_H    0.0488362 0.0036879   13.242  < 2e-16 ***
TEAM_FIELDING_E -0.0191370  0.0023989  -7.978 2.35e-15 ***
TEAM_BASERUN_SB   0.0249146  0.0042097   5.918 3.75e-09 ***
TEAM_FIELDING_DP -0.1219441  0.0129328  -9.429  < 2e-16 ***
TEAM_PITCHING_HR  0.0149619  0.0210392   0.711 0.477067
TEAM_BATTING_3B   0.0649257  0.0167861   3.868 0.000113 ***
TEAM_BATTING_BB   0.0115657  0.0033753   3.427 0.000622 ***
TEAM_BATTING_2B  -0.0211889  0.0091673  -2.311 0.020902 *
TEAM_PITCHING_SO  0.0029431  0.0006728   4.375 1.27e-05 ***
TEAM_BATTING_SO  -0.0085618  0.0024539  -3.489 0.000494 ***
TEAM_PITCHING_H  -0.0008248  0.0003279  -2.515 0.011967 *
TEAM_BATTING_HR   0.0517627  0.0240121   2.156 0.031213 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.07 on 2263 degrees of freedom
Multiple R-squared:  0.3152, Adjusted R-squared:  0.3116
F-statistic: 86.81 on 12 and 2263 DF,  p-value: < 2.2e-16
```

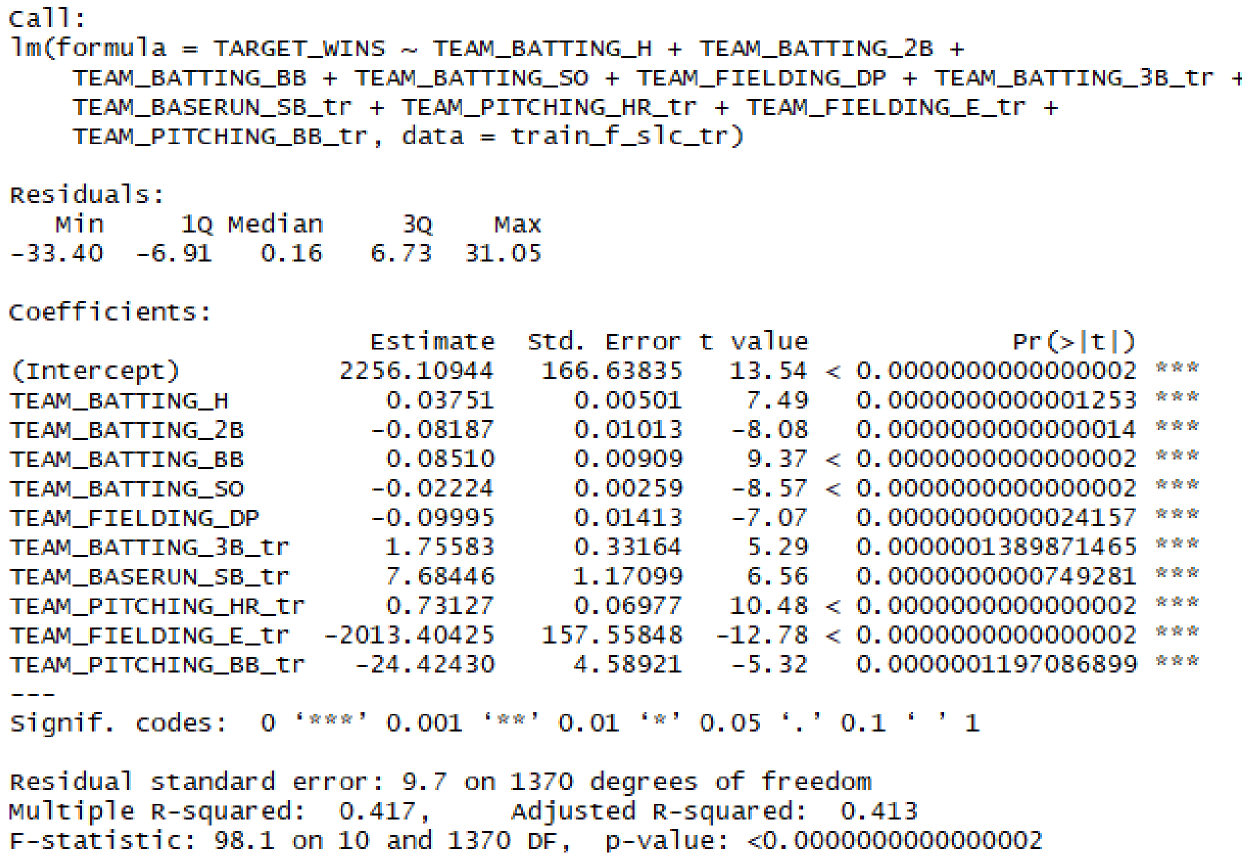


Similar to model-1, model-2 also have 12 predictors. Again, we see the counterintuitive problem. The coefficients for all the predictors but TEAM_BATTING_2B make sense. TEAM_BATTING_2B coefficient did not represent that this predictor having positive effects on winning but was negative.

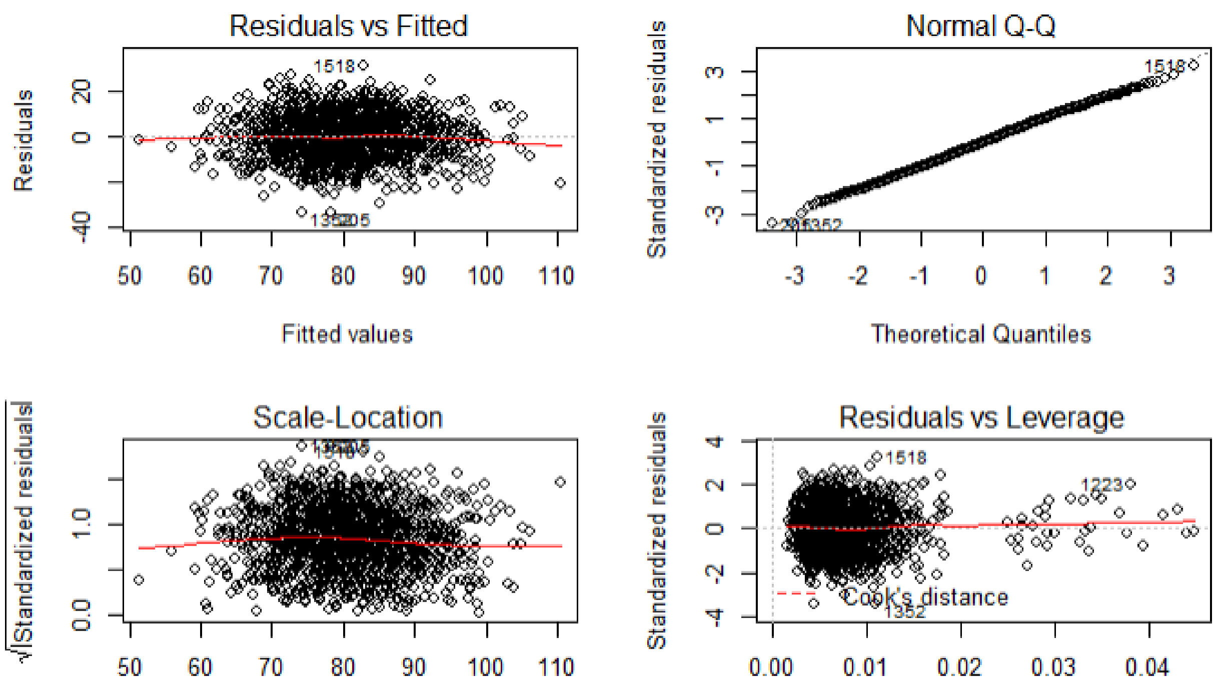
The diagnostic plots indicated that all the assumptions are valid. The residuals have constant variance. No outliers beyond the Cook's distance that could have strong influence on the regression results.

Model-3. Box-Cox Transformation + Backward Elimination

The 3rd model is created using the backward elimination process on the variables transformed using box cox transformation. A summary of model-3 is as follows:



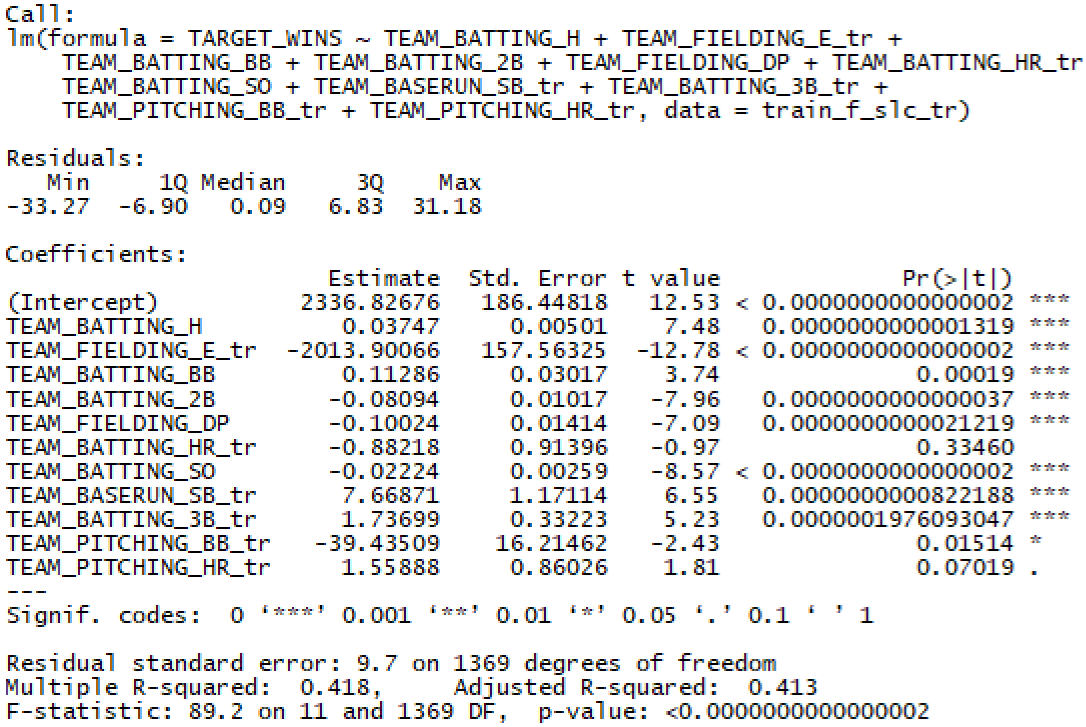
```
Call:
lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + 
    TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_FIELDING_DP + TEAM_BATTING_3B_tr + 
    TEAM_BASERUN_SB_tr + TEAM_PITCHING_HR_tr + TEAM_FIELDING_E_tr + 
    TEAM_PITCHING_BB_tr, data = train_f_slc_tr)

Residuals:
   Min     1Q Median     3Q    Max 
-33.40  -6.91   0.16   6.73  31.05 

Coefficients:
                      Estimate  Std. Error t value             Pr(>|t|)    
(Intercept)         2256.10944   166.63835   13.54 < 0.0000000000000002 ***
TEAM_BATTING_H         0.03751     0.00501    7.49   0.0000000000001253 ***
TEAM_BATTING_2B       -0.08187     0.01013   -8.08   0.0000000000000014 ***
TEAM_BATTING_BB        0.08510     0.00909    9.37 < 0.0000000000000002 ***
TEAM_BATTING_SO       -0.02224     0.00259   -8.57 < 0.0000000000000002 ***
TEAM_FIELDING_DP      -0.09995     0.01413   -7.07   0.0000000000024157 ***
TEAM_BATTING_3B_tr     1.75583     0.33164    5.29   0.0000001389871465 ***
TEAM_BASERUN_SB_tr     7.68446     1.17099    6.56   0.0000000000749281 ***
TEAM_PITCHING_HR_tr    0.73127     0.06977   10.48 < 0.0000000000000002 ***
TEAM_FIELDING_E_tr -2013.40425   157.55848  -12.78 < 0.0000000000000002 ***
TEAM_PITCHING_BB_tr  -24.42430     4.58921   -5.32   0.0000001197086899 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.7 on 1370 degrees of freedom
Multiple R-squared:  0.417,	Adjusted R-squared:  0.413 
F-statistic: 98.1 on 10 and 1370 DF,  p-value: <0.0000000000000002
```



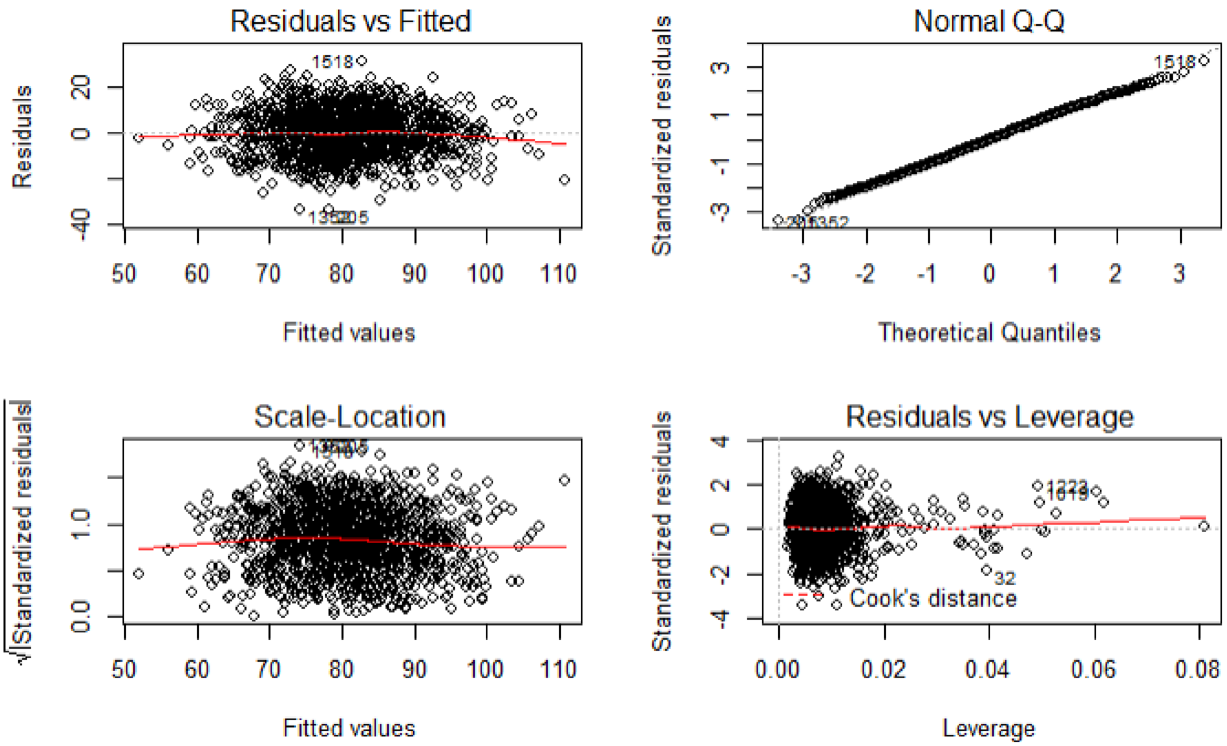
This model has less predictors (10) compared to the previous two. Again, we saw a negative TEAM_BATTING_2B coefficient, which made no sense. In addition, TEAM_PITCHING_HR_tr is expected to be negative as the pitching team HR will increase the chase of the losing of the batting team. The diagnostic plots indicated that all the assumptions are valid. The residuals have constant variance. No outliers beyond the Cook's distance that could have strong influence on the regression results.

Model-4. Box-Cox Transformation + Forward Stepwise Selection

The 4th model is created using the forward selection process. The full model used in this process is the same as the one used in model-1. A null model started from a null model and searched through models lying in the range between the null and full model using the forward selection algorithm. The resulting model with variables was selected as our final model-2. According to this procedure, the best model is the one that includes the variables TEAM_BATTING_H , TEAM_BATTING_2B , TEAM_BATTING_BB , TEAM_BATTING_SO , TEAM_FIELDING_DP,TEAM_BATTING_3B_tr, TEAM_BATTING_HR_tr, TEAM_BASERUN_SB_tr, TEAM_FIELDING_E_tr.. So the resulting model of forward stepwise selection is the same as the one resulted from the backward stepwise selection. A summary of model-4 is as follows:



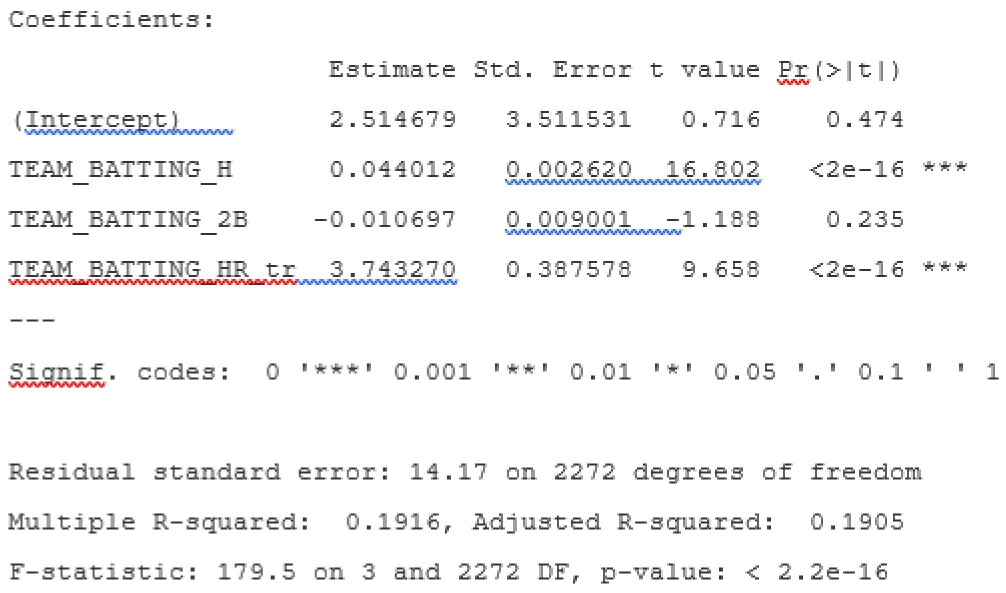
```
Call:
lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_FIELDING_E_tr + 
    TEAM_BATTING_BB + TEAM_BATTING_2B + TEAM_FIELDING_DP + TEAM_BATTING_HR_tr + 
    TEAM_BATTING_SO + TEAM_BASERUN_SB_tr + TEAM_BATTING_3B_tr + 
    TEAM_PITCHING_BB_tr + TEAM_PITCHING_HR_tr, data = train_f_slc_tr)

Residuals:
   Min     1Q Median     3Q    Max 
-33.27  -6.90   0.09   6.83  31.18 

Coefficients:
                       Estimate  Std. Error t value             Pr(>|t|)    
(Intercept)          2336.82676   186.44818   12.53 < 0.0000000000000002 ***
TEAM_BATTING_H          0.03747     0.00501    7.48   0.0000000000001319 ***
TEAM_FIELDING_E_tr  -2013.90066   157.56325  -12.78 < 0.0000000000000002 ***
TEAM_BATTING_BB         0.11286     0.03017    3.74              0.00019 ***
TEAM_BATTING_2B        -0.08094     0.01017   -7.96   0.0000000000000037 ***
TEAM_FIELDING_DP       -0.10024     0.01414   -7.09   0.0000000000021219 ***
TEAM_BATTING_HR_tr     -0.88218     0.91396   -0.97              0.33460    
TEAM_BATTING_SO        -0.02224     0.00259   -8.57 < 0.0000000000000002 ***
TEAM_BASERUN_SB_tr      7.66871     1.17114    6.55   0.0000000000822188 ***
TEAM_BATTING_3B_tr      1.73699     0.33223    5.23   0.0000001976093047 ***
TEAM_PITCHING_BB_tr   -39.43509    16.21462   -2.43              0.01514 *  
TEAM_PITCHING_HR_tr     1.55888     0.86026    1.81              0.07019 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.7 on 1369 degrees of freedom
Multiple R-squared:  0.418,	Adjusted R-squared:  0.413 
F-statistic: 89.2 on 11 and 1369 DF,  p-value: <0.0000000000000002
```



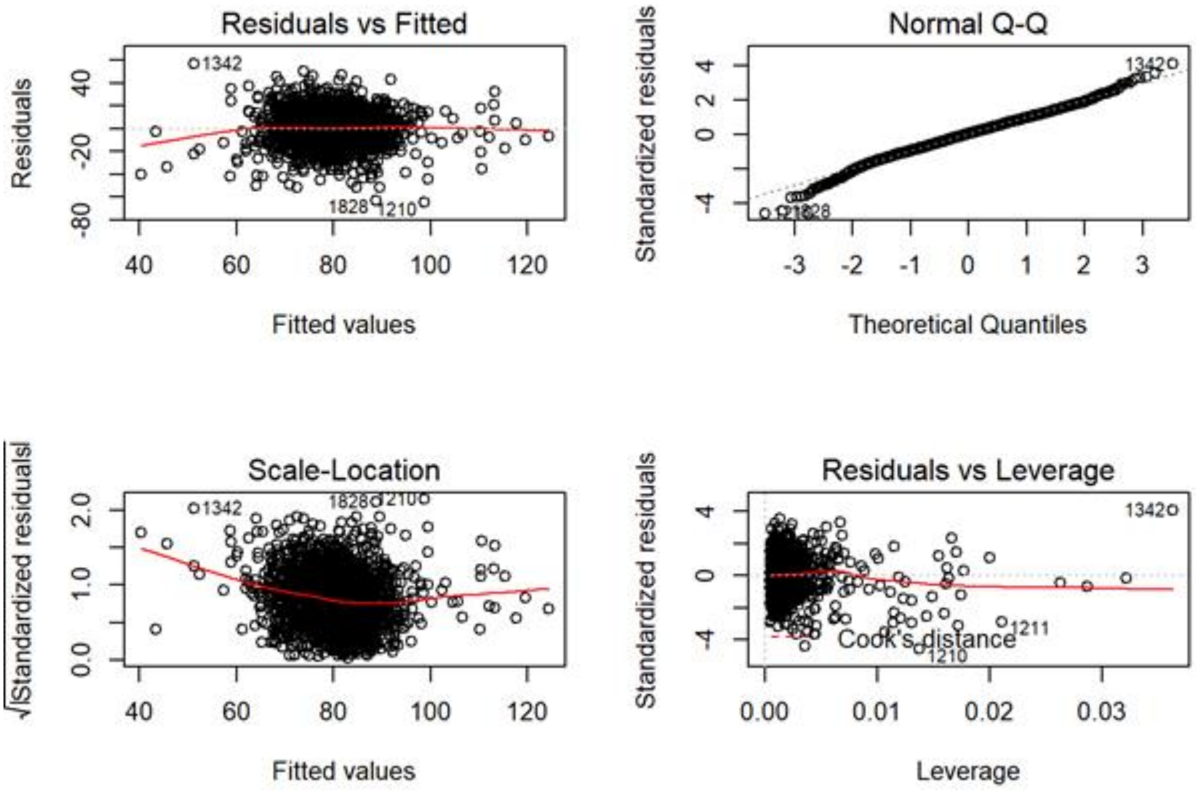
The coefficients of model-4 are very close to model-3, suggesting backward and forward stepwise selection reached similar results. Similarly, we saw negative coefficients for TEAM_BATTING_2B and TEAM_PITCHING_HR_tr, which made no sense. The diagnostic plots indicated that all the assumptions are valid. The residuals have constant variance. No outliers beyond the Cook's distance that could have strong influence on the regression results.

Model-5. Manually Selection - Strategy I

Model-5 was simple regression model built by selecting the variables that were correlated with the dependent variable. The correlation matrix was used to select the correlation of the independent variables with dependent variables to select the variables for our model. The transformed variables using log transformation were included in the model. Based on the correlation matrix, we selected TEAM_BATTING_H, TEAM_BATTING_2B, TEAM_BATTING_HR_tr as our independent variables for prediction. Summary for model-5:



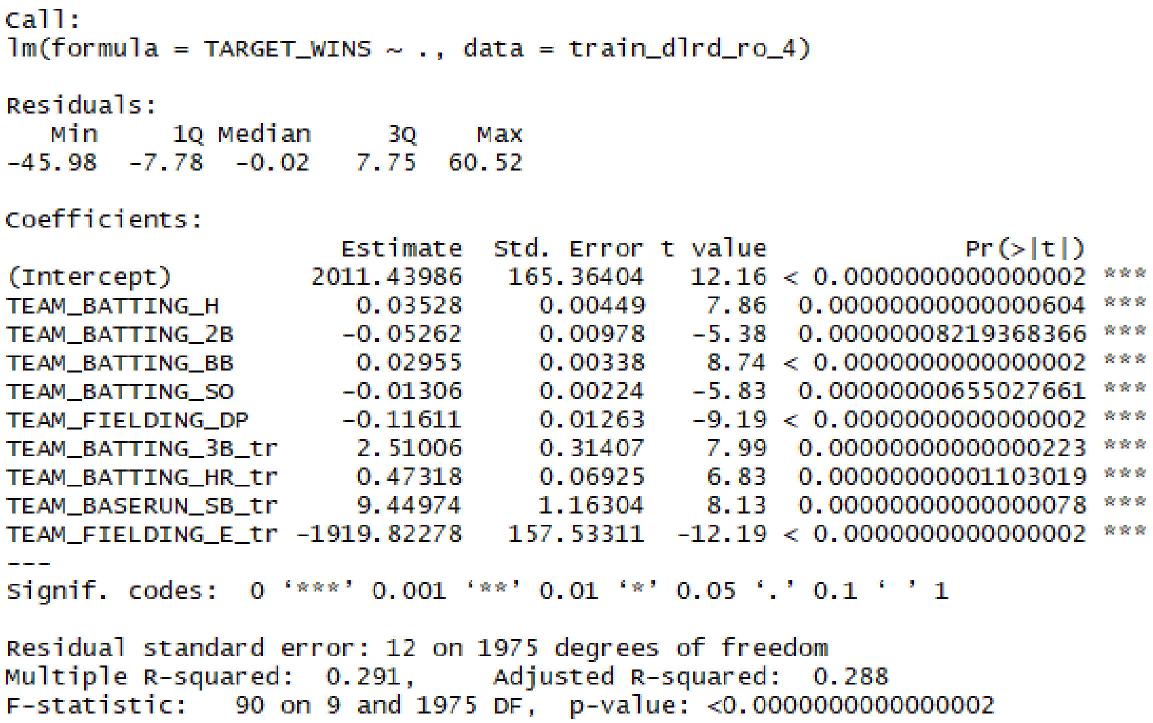
```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.514679   3.511531   0.716    0.474
TEAM_BATTING_H      0.044012   0.002620  16.802   <2e-16 ***
TEAM_BATTING_2B    -0.010697   0.009001  -1.188    0.235
TEAM_BATTING_HR_tr  3.743270   0.387578   9.658   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.17 on 2272 degrees of freedom
Multiple R-squared:  0.1916, Adjusted R-squared:  0.1905
F-statistic: 179.5 on 3 and 2272 DF, p-value: < 2.2e-16
```



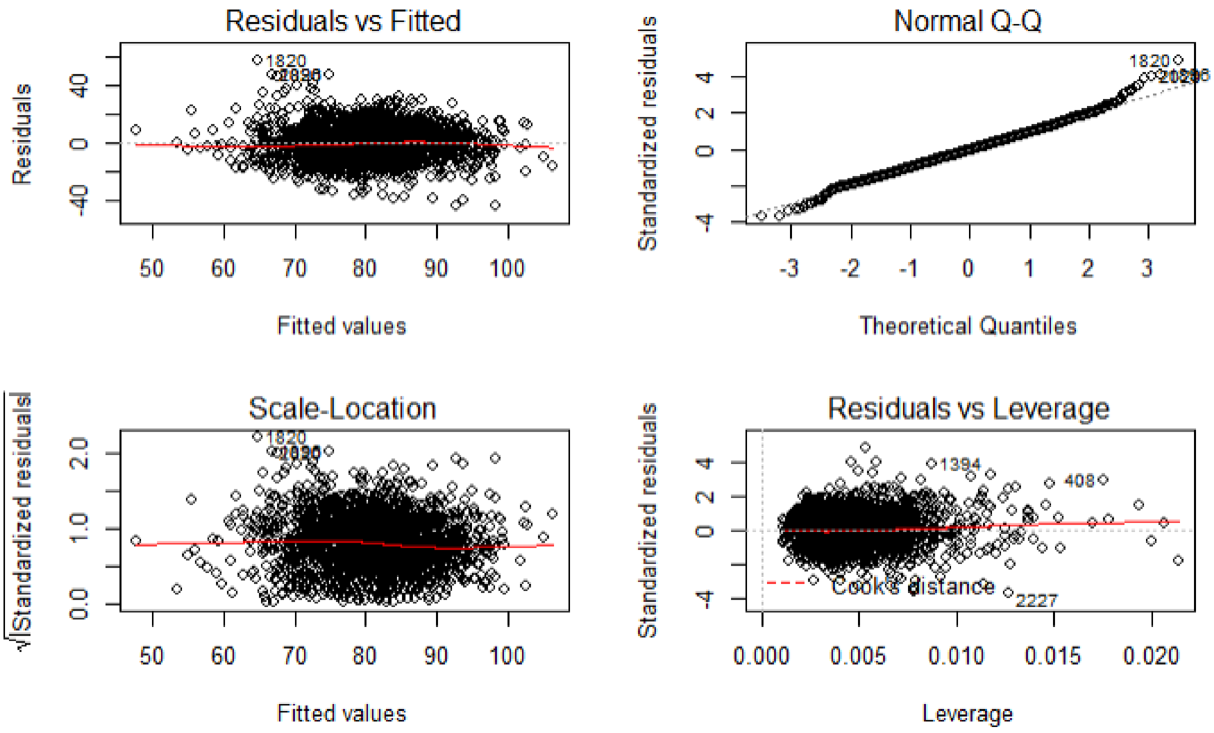
Manually selection resulted a model which is different from those selected by computer. It had much less predictor variables. We saw negative coefficients for TEAM_BATTING_2B again. The coefficients for the other predictors made sense. The diagnostic plots indicated that all the assumptions are valid. The residuals have constant variance. No outliers beyond the Cook's distance that could have strong influence on the regression results.

Model-6. Manually Selection - Strategy II

In model-6, we selected the variables before fitting the linear regression model by the consequent steps: removing highly correlated predictor variables, imputing the missing values, transforming the skewed variables and removing outlier. A summary of model-6 is as follows:



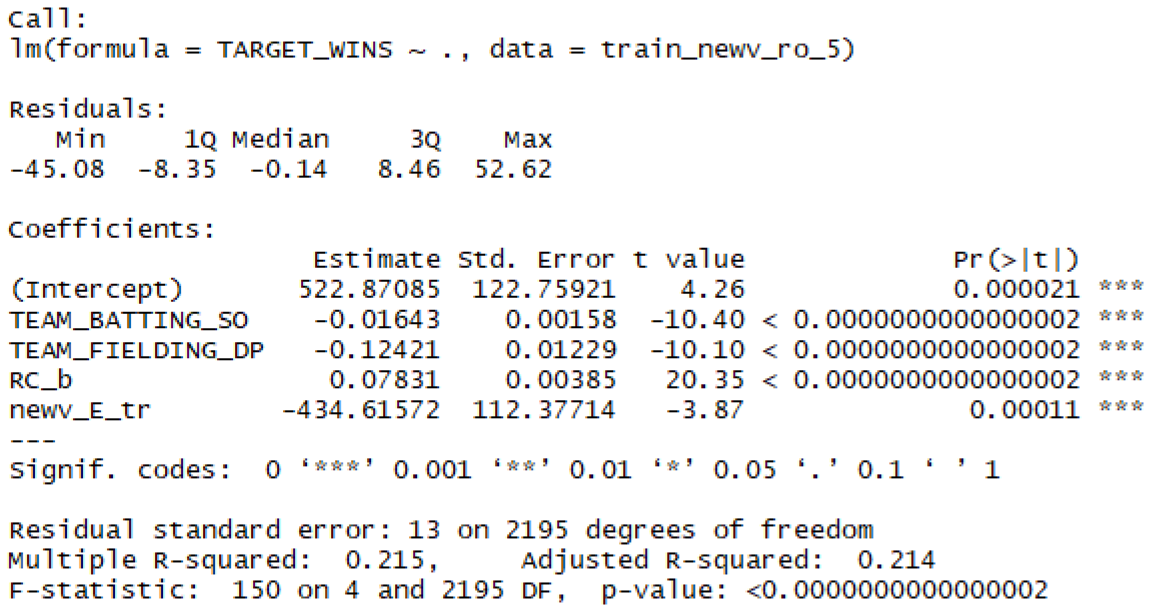
```
Call:
lm(formula = TARGET_WINS ~ ., data = train_dlrd_ro_4)

Residuals:
   Min     1Q Median     3Q    Max 
-45.98  -7.78  -0.02   7.75  60.52 

Coefficients:
                      Estimate  Std. Error t value             Pr(>|t|)    
(Intercept)         2011.43986   165.36404   12.16 < 0.0000000000000002 ***
TEAM_BATTING_H         0.03528     0.00449    7.86   0.0000000000000604 ***
TEAM_BATTING_2B       -0.05262     0.00978   -5.38   0.0000000821936836 ***
TEAM_BATTING_BB        0.02955     0.00338    8.74 < 0.0000000000000002 ***
TEAM_BATTING_SO       -0.01306     0.00224   -5.83   0.0000000065502766 ***
TEAM_FIELDING_DP      -0.11611     0.01263   -9.19 < 0.0000000000000002 ***
TEAM_BATTING_3B_tr     2.51006     0.31407    7.99   0.0000000000000223 ***
TEAM_BATTING_HR_tr     0.47318     0.06925    6.83   0.0000000000110302 ***
TEAM_BASERUN_SB_tr     9.44974     1.16304    8.13   0.0000000000000078 ***
TEAM_FIELDING_E_tr -1919.82278   157.53311  -12.19 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12 on 1975 degrees of freedom
Multiple R-squared:  0.291,     Adjusted R-squared:  0.288 
F-statistic:   90 on 9 and 1975 DF,  p-value: <0.0000000000000002
```



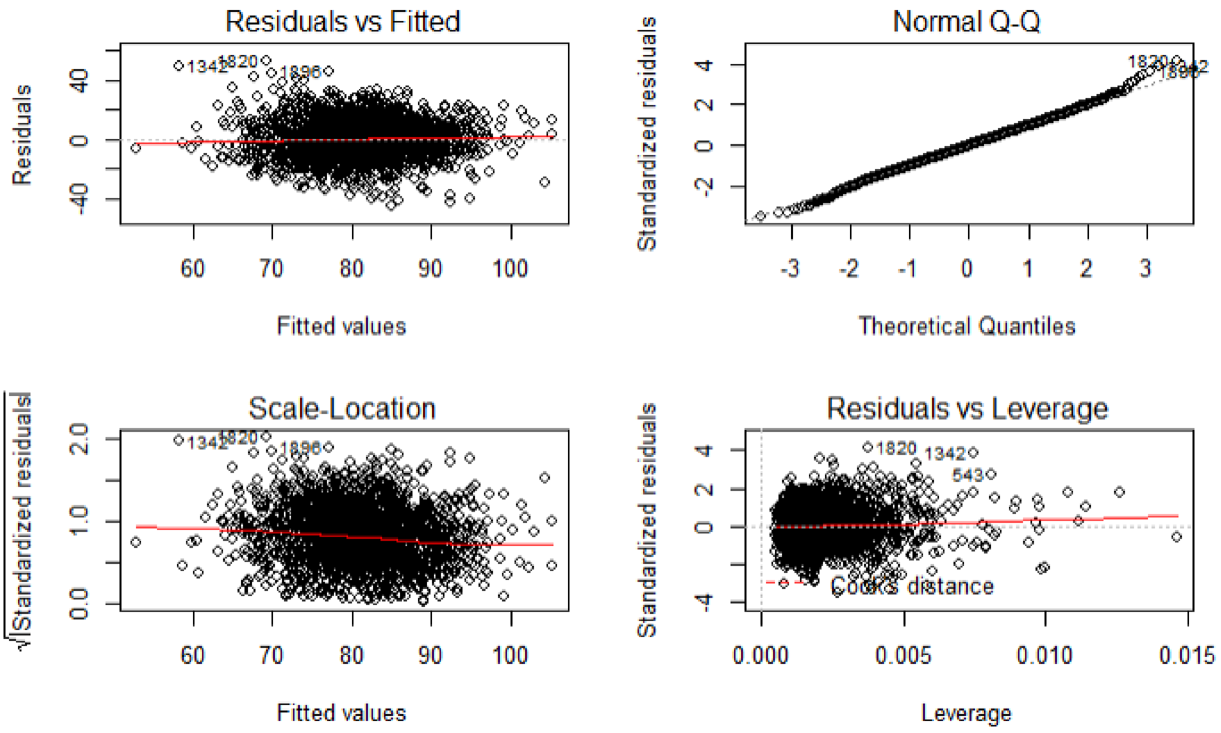
Manually selection of predictors with a different strategy resulted a model which is different from the one reached by picking from the correlation matrix. It had relative more predictor variables. We saw negative coefficients for TEAM_BATTING_2B again. The coefficients for the other predictors made sense. The diagnostic plots indicated that all the assumptions are valid. The residuals have constant variance. No outliers beyond the Cook's distance that could have strong influence on the regression results.

Model-7. Basic Runs Created

In this model we generate a new variable "runs created (RC_b)" by putting variables into one formula to calculate the chance of scoring. Because of the introduction of this new variables we deleted the five variables (TEAM_BATTING_H, TEAM_BATTING_2B, TEAM_BATTING_3B, TEAM_BATTING_HR and TEAM_BATTING_BB) involved in the formula. The new variable RC_b is normal distributed and with no skewness issue as some of the components. A summary of model-7 is as below:



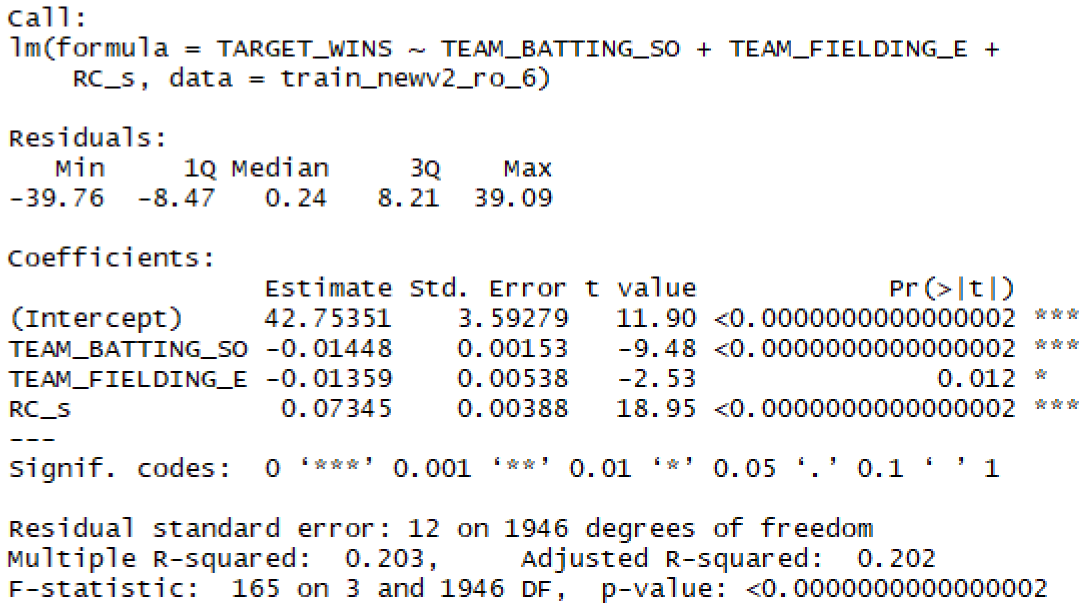
```
Call:
lm(formula = TARGET_WINS ~ ., data = train_newv_ro_5)

Residuals:
   Min     1Q Median     3Q    Max 
-45.08  -8.35  -0.14   8.46  52.62 

Coefficients:
                   Estimate Std. Error t value             Pr(>|t|)    
(Intercept)       522.87085  122.75921    4.26             0.000021 ***
TEAM_BATTING_SO    -0.01643    0.00158  -10.40 < 0.0000000000000002 ***
TEAM_FIELDING_DP   -0.12421    0.01229  -10.10 < 0.0000000000000002 ***
RC_b                0.07831    0.00385   20.35 < 0.0000000000000002 ***
newv_E_tr        -434.61572  112.37714   -3.87              0.00011 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13 on 2195 degrees of freedom
Multiple R-squared:  0.215,	Adjusted R-squared:  0.214 
F-statistic:  150 on 4 and 2195 DF,  p-value: <0.0000000000000002
```



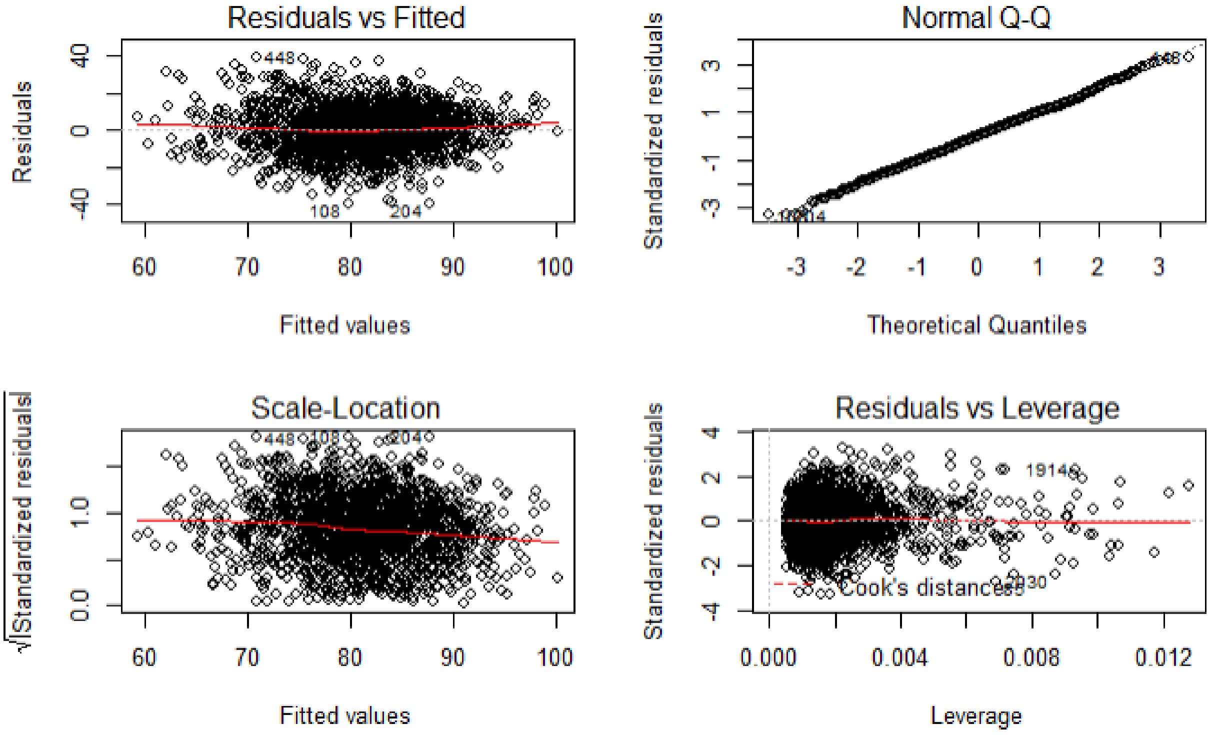
The introduction of the new predictors by combining some predictors led to the dramatic decrease of the number of variables. The coefficients for TEAM_BATTING_SO and TEAM_FIELDING_DP made sense. The coefficients for the Run Created was negative, matching its correlation with winning. The diagnostic plots indicated that all the assumptions are valid. The residuals have constant variance. No outliers beyond the Cook's distance that could have strong influence on the regression results.

Model-8. "Stolen Based" Runs Created

The approach for our 8th model is similar to that of Model-6. We generate a stolen based version of 'runs created' (RC_sc). Two more(TEAM_BASERUN_SB and TEAM_BASERUN_CS) variables were included in the new variable so we could further eliminate these two variables. The new variable RC_b is normal distributed. A summary of model-8 is as below:



```
Call:
lm(formula = TARGET_WINS ~ TEAM_BATTING_SO + TEAM_FIELDING_E +
    RC_s, data = train_newv2_ro_6)

Residuals:
   Min     1Q Median     3Q    Max 
-39.76  -8.47   0.24   8.21  39.09 

Coefficients:
                Estimate Std. Error t value             Pr(>|t|)    
(Intercept)     42.75351    3.59279   11.90 <0.0000000000000002 ***
TEAM_BATTING_SO -0.01448    0.00153   -9.48 <0.0000000000000002 ***
TEAM_FIELDING_E -0.01359    0.00538   -2.53                0.012 *  
RC_s             0.07345    0.00388   18.95 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12 on 1946 degrees of freedom
Multiple R-squared:  0.203,	Adjusted R-squared:  0.202 
F-statistic:  165 on 3 and 1946 DF,  p-value: <0.0000000000000002
```

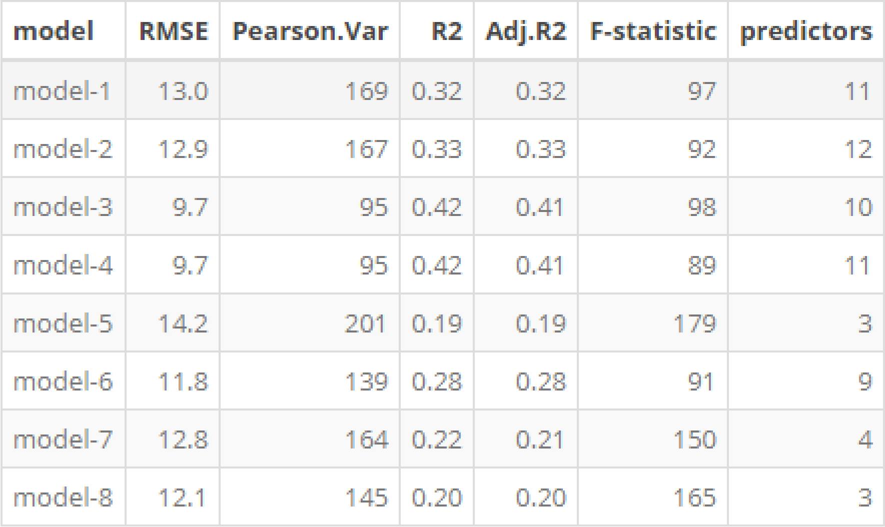


As a similar version to model-7, the coefficients for TEAM_BATTING_SO and TEAM_FIELDING_DP made sense in this model. The coefficients for the Run Created was negative, matching its correlation with winning. The diagnostic plots indicated that all the assumptions are valid. The residuals have constant variance. No outliers beyond the Cook's distance that could have strong influence on the regression results.

4. SELECT MODELS

4.1 Compare the Key Statistic

First, we compared the statistic for different models as shown in the table below. Although Model-3 did not have the lowest RMSE, it was the most parsimonious (fewest variables) and stable (little collinearity between variables).



model	RMSE	Pearson.Var	R2	Adj.R2	F-statistic	predictors
model-1	13.0	169	0.32	0.32	97	11
model-2	12.9	167	0.33	0.33	92	12
model-3	9.7	95	0.42	0.41	98	10
model-4	9.7	95	0.42	0.41	89	11
model-5	14.2	201	0.19	0.19	179	3
model-6	11.8	139	0.28	0.28	91	9
model-7	12.8	164	0.22	0.21	150	4
model-8	12.1	145	0.20	0.20	165	3

We know the higher the R-squared, the stronger the correlation between the predictors and response variable. But the R-square is less important when the interest is not in prediction rather than in the relationship between variables. When we look at RMSE, lower values of RMSE indicate better fit. As for the the F-test statistic, a high the F-test value suggests the proposed relationship between the response variable and the set of predictors is more statistically reliable. Obviously model-3 is the best as it has the lowest RMSE and Person estimated residual variance, highest R-square and F-statistic value. Based on the key statistic (RMSE-R2-F), the models from best to the worst are: model-3, model-4, model-6, model-8, model-7, model-2, model-1, model-5.

4.2 Compare the Residuals Behaviors

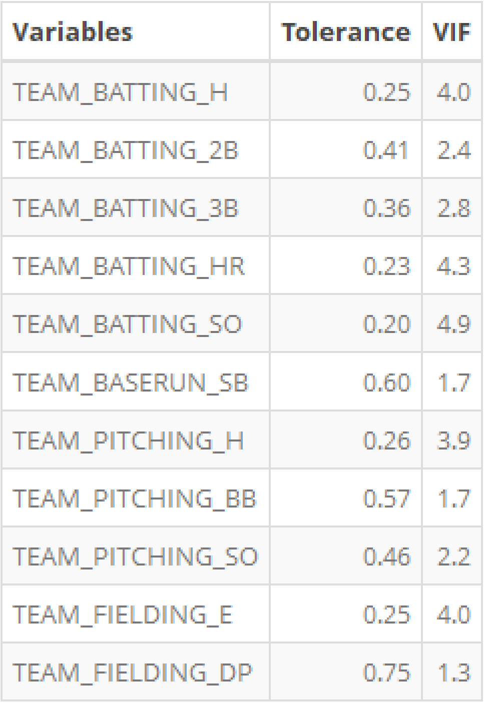
We then checked back the diagnostic plots for each model to compare the residuals behaviors to see how well each model represented the data. From the Residuals vs Fitted (upper left panel) and Scale-Location(lower left panel), we could see that the residuals of all the models show constant variance. From the Q-Q plot, we see that the residuals of all the models follow normal distribution. The Residuals vs Leverage plot (lower right) showed that there is no cases outside of the Cook’s distance which could be very influential and cause the unstable of the model. In other words, all the models passed the quality control.

4.3 Check the Multicollinearity

Then we go to check whether multicollinearity presented in each model. VIF for each predictor was measured for each model. The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

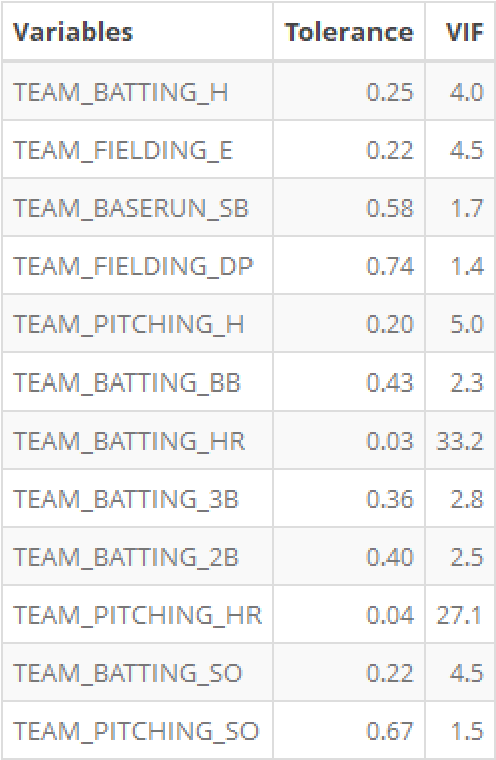
As shown in the following tables, model-3, model-6, model-8, model-7, model-5 have no predictor variable VIF>10.

Model-1 Variance inflation factors measure



Variables	Tolerance	VIF
TEAM_BATTING_H	0.25	4.0
TEAM_BATTING_2B	0.41	2.4
TEAM_BATTING_3B	0.36	2.8
TEAM_BATTING_HR	0.23	4.3
TEAM_BATTING_SO	0.20	4.9
TEAM_BASERUN_SB	0.60	1.7
TEAM_PITCHING_H	0.26	3.9
TEAM_PITCHING_BB	0.57	1.7
TEAM_PITCHING_SO	0.46	2.2
TEAM_FIELDING_E	0.25	4.0
TEAM_FIELDING_DP	0.75	1.3

Model-2 Variance inflation factors measure



Variables	Tolerance	VIF
TEAM_BATTING_H	0.25	4.0
TEAM_FIELDING_E	0.22	4.5
TEAM_BASERUN_SB	0.58	1.7
TEAM_FIELDING_DP	0.74	1.4
TEAM_PITCHING_H	0.20	5.0
TEAM_BATTING_BB	0.43	2.3
TEAM_BATTING_HR	0.03	33.2
TEAM_BATTING_3B	0.36	2.8
TEAM_BATTING_2B	0.40	2.5
TEAM_PITCHING_HR	0.04	27.1
TEAM_BATTING_SO	0.22	4.5
TEAM_PITCHING_SO	0.67	1.5

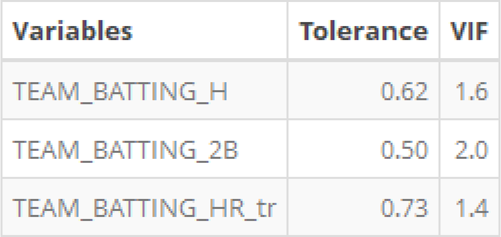
Model-3 Variance inflation factors measure

Variables	Tolerance	VIF
TEAM_BATTING_H	0.27	3.8
TEAM_BATTING_2B	0.40	2.5
TEAM_BATTING_BB	0.14	7.0
TEAM_BATTING_SO	0.27	3.7
TEAM_FIELDING_DP	0.91	1.1
TEAM_BATTING_3B_tr	0.45	2.2
TEAM_BASERUN_SB_tr	0.80	1.2
TEAM_PITCHING_HR_tr	0.34	3.0
TEAM_FIELDING_E_tr	0.52	1.9
TEAM_PITCHING_BB_tr	0.14	7.2

Model-4 Variance inflation factors measure

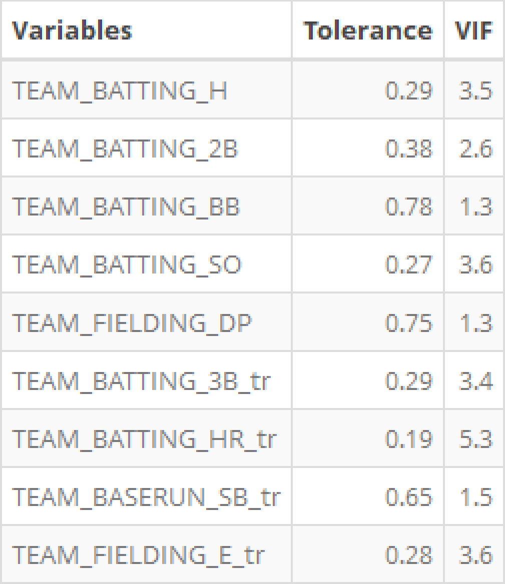
Variables	Tolerance	VIF
TEAM_BATTING_H	0.27	3.8
TEAM_FIELDING_E_tr	0.52	1.9
TEAM_BATTING_BB	0.01	76.6
TEAM_BATTING_2B	0.40	2.5
TEAM_FIELDING_DP	0.91	1.1
TEAM_BATTING_HR_tr	0.00	465.5
TEAM_BATTING_SO	0.27	3.7
TEAM_BASERUN_SB_tr	0.80	1.2
TEAM_BATTING_3B_tr	0.45	2.2
TEAM_PITCHING_BB_tr	0.01	90.2
TEAM_PITCHING_HR_tr	0.00	453.2

Model-5 Variance inflation factors measure



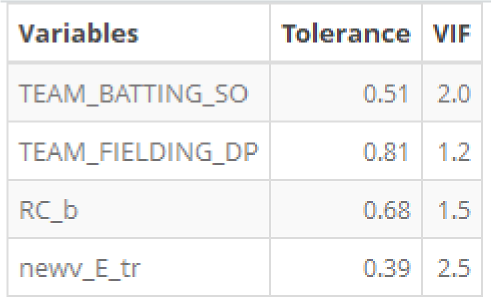
Variables	Tolerance	VIF
TEAM_BATTING_H	0.62	1.6
TEAM_BATTING_2B	0.50	2.0
TEAM_BATTING_HR_tr	0.73	1.4

Model-6 Variance inflation factors measure



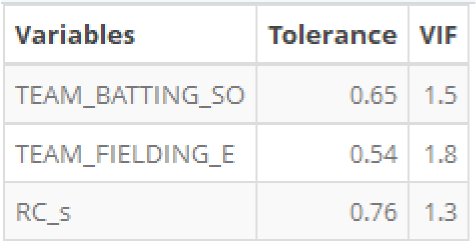
Variables	Tolerance	VIF
TEAM_BATTING_H	0.29	3.5
TEAM_BATTING_2B	0.38	2.6
TEAM_BATTING_BB	0.78	1.3
TEAM_BATTING_SO	0.27	3.6
TEAM_FIELDING_DP	0.75	1.3
TEAM_BATTING_3B_tr	0.29	3.4
TEAM_BATTING_HR_tr	0.19	5.3
TEAM_BASERUN_SB_tr	0.65	1.5
TEAM_FIELDING_E_tr	0.28	3.6

Model-7 Variance inflation factors measure



Variables	Tolerance	VIF
TEAM_BATTING_SO	0.51	2.0
TEAM_FIELDING_DP	0.81	1.2
RC_b	0.68	1.5
newv_E_tr	0.39	2.5

Model-8 Variance inflation factors measure



Variables	Tolerance	VIF
TEAM_BATTING_SO	0.65	1.5
TEAM_FIELDING_E	0.54	1.8
RC_s	0.76	1.3

4.4 Compare the Number of Predictors

Then we compare the number of number of the predictors in the left five models. The model has the fewest variables is the most parsimonious. Model-3 and model-6 each has more than 10 predictors while model-8, model-7 and model-5 has less than 4 predictors.

Not only model-3 and 6 are less favorable because they have much more predictors than the other models, one of the coefficients of these two models suggested they may not be good models. TEAM_BATTING_2B should have positive influence to the target variable TARGET_WINS. But the estimated coefficient for TEAM_BATTING_2B in model-3 and model-6 are both negative. This make no sense. Model-5 has the same issue too.

Last, we compare the F-statistic for the rest 2 models and found that is slightly better than model-8. The value F-statistic value, 179, of model-7 is also the highest in all the models, indicating the proposed relationship between the response variable and the set of predictors in model-7 is statistically reliable. Although R-squared of model-7 did not show a good possible fit and RMSE was not the lowest in all models, we still think model-7 fit our need. Because our interest is not in the relationship between variables but in prediction, so the R-square is less important. In conclusion, model-7 is the most stable and parsimonious model. Therefore, **we select the model-7 (Basic Runs Created)** as the model for the prediction of TARGET_WINS with the evaluation data set. The TARGET_WINS could be predict with this formula:

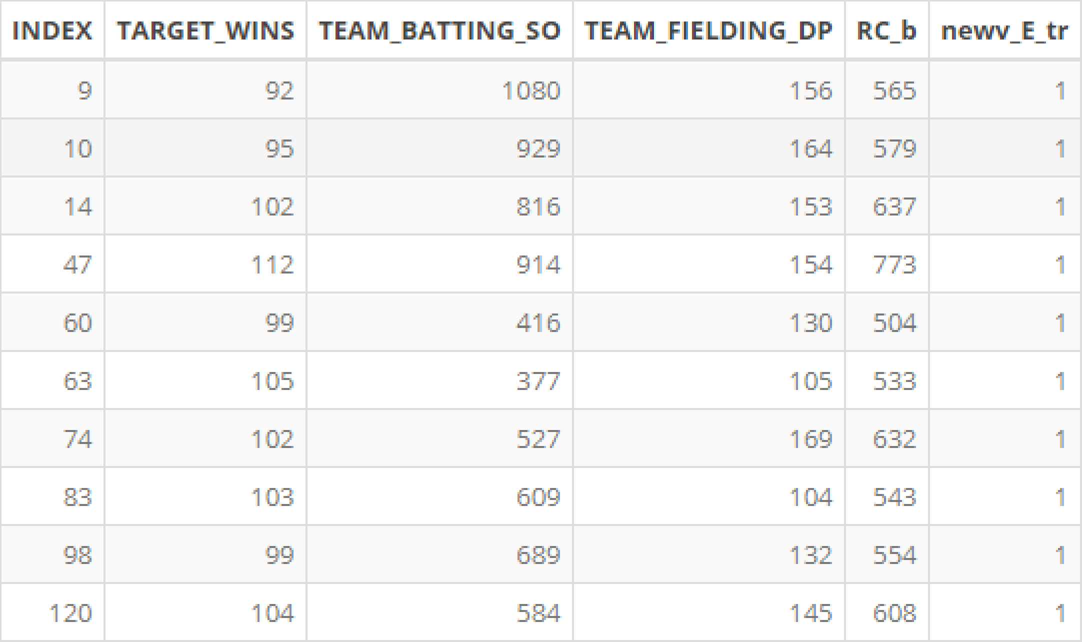
TARGET_WINS = 522.87-0.0164*TEAM_BATTING_SO-0.1242*TEAM_FIELDING_DP+0.0783*RC_B-434.62*new_E_tr

4.5 Test Evaluation Data

To test the efficacy of the model, we make prediction using the valuation data set. Before apply the model on the data set, we treated the evaluation data set with the same process as to the train data set: eliminating collinear predictors, imputation, transformation, outlier removal. The whole predicted data set could be found through this URL:

https://raw.githubusercontent.com/YunMai-SPS/DATA621_homework/master/data621_assignment1/moneyball-evaluation-pred.csv

The first 10 rows of that predicted results is displayed as an example:



INDEX	TARGET_WINS	TEAM_BATTING_SO	TEAM_FIELDING_DP	RC_b	newv_E_tr
9	92	1080	156	565	1
10	95	929	164	579	1
14	102	816	153	637	1
47	112	914	154	773	1
60	99	416	130	504	1
63	105	377	105	533	1
74	102	527	169	632	1
83	103	609	104	543	1
98	99	689	132	554	1
120	104	584	145	608	1

5. REFERENCES:

https://www.statmethods.net/stats/descriptives.html

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4842399/

http://flowingdata.com/2012/05/15/how-to-visualize-and-compare-distributions/

Wikipedia: Run (baseball)

https://en.wikipedia.org/wiki/Run_(baseball)

Wikipedia: Runs created

https://en.wikipedia.org/wiki/Runs_created

Wikipedia: At bat

https://en.wikipedia.org/wiki/At_bat

Wikipedia: Batting average

https://en.wikipedia.org/wiki/Batting_average

Evaluating Batting

http://baseballstats.tripod.com/batting.html

Appendix

RCODE:

Github: https://github.com/YunMai-SPS/DATA621_homework/blob/master/data621_assignment1/DATA621_Assignment_1_combine.Rmd