Home Work Assignment - 01

Critical Thinking Group 5

Contents

Overview

The data set contains approximately 2200 records. Each record represents a professional baseball team from the years 1871 to 2006 inclusive. Each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season. We will be exploring, analyzing, and modeling the data set to predict a number of wins for a team using Ordinary Least Square (OLS).
To attain our objective, we will be following the below best practice steps and guidelines:
1 -Data Exploration
2 -Data Preparation
3 -Build Models
4 -Select Models

1 Data Exploration Analysis

In section we will explore and gain some insights into the dataset by pursuing the below high level steps and inquiries:
-Variable identification
-Variable Relationships
-Data summary analysis
-Outliers and Missing Values Identification

1.1 Variable identification

First let's display and examine the data dictionary or the data columns as shown in table 1.

Table 1: Variable Definition

VARIABLE_NAME	DEFINITION	THEORETICAL_EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	Target
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

We notice that all variables are numeric. The variable names seem to follow certain naming pattern to highlight certain arithmetic relationships. In other words, we can compute the number of '1B' hits by taking the difference between overall hits and '2B', '3B', 'HR'. Although such naming and construct is not recommended in normalized database design (as it violates third normal form), it is very frequent practice in the data analytics.

Our predictor input is made of 15 variables. And our dependent variable is one variable called TARGET_WINS.

Please note that we will not be using INDEX variable as it serves as just an identifier for each row. And has no relationships to other variables.

1.2 Data Summary Analysis

In this section, we will create summary data to better understand the initial relationship variables have with our dependent variable using correlation, central tendency, and dispersion As shown in table 2.

Table 2: Data Summary

	mean	sd	median	trimmed
TARGET_WINS	80.79086	15.75215	82.0	81.31229
TEAM_BATTING_H	1469.26977	144.59120	1454.0	1459.04116
TEAM_BATTING_2B	241.24692	46.80141	238.0	240.39627
TEAM_BATTING_3B	55.25000	27.93856	47.0	52.17563
TEAM_BATTING_HR	99.61204	60.54687	102.0	97.38529
TEAM_BATTING_BB	501.55888	122.67086	512.0	512.18331
TEAM_BATTING_SO	735.60534	248.52642	750.0	742.31322
TEAM_BASERUN_SB	124.76177	87.79117	101.0	110.81188
TEAM_BASERUN_CS	52.80386	22.95634	49.0	50.35963
TEAM_BATTING_HBP	59.35602	12.96712	58.0	58.86275
TEAM_PITCHING_H	1779.21046	1406.84293	1518.0	1555.89517
TEAM_PITCHING_HR	105.69859	61.29875	107.0	103.15697
TEAM_PITCHING_BB	553.00791	166.35736	536.5	542.62459
TEAM_PITCHING_SO	817.73045	553.08503	813.5	796.93391
TEAM_FIELDING_E	246.48067	227.77097	159.0	193.43798
TEAM_FIELDING_DP	146.38794	26.22639	149.0	147.57789

Table 3: Missing Data and Data Correlation

	Missing	Correlation
TARGET_WINS	0	1.0000000
TEAM_BATTING_H	0	0.3887675
TEAM_BATTING_2B	0	0.2891036
TEAM_BATTING_3B	0	0.1426084
TEAM_BATTING_HR	0	0.1761532
TEAM_BATTING_BB	0	0.2325599
TEAM_BATTING_SO	102	-0.0317507
TEAM_BASERUN_SB	131	0.1351389
TEAM_BASERUN_CS	772	0.0224041
TEAM_BATTING_HBP	2085	0.0735042
TEAM_PITCHING_H	0	-0.1099371
TEAM_PITCHING_HR	0	0.1890137
TEAM_PITCHING_BB	0	0.1241745
TEAM_PITCHING_SO	102	-0.0784361
TEAM_FIELDING_E	0	-0.1764848
TEAM_FIELDING_DP	286	-0.0348506

Based on table 2 and Table 3, we can make the below observations:

1.Some of the variables like TEAM_PITCHING_H, TEAM_PITCHING_SO and TEAM_FIELDING_E seem to have outliers which is evident from the mean, median and trimmed mean values.

2.TEAM_BATTING_HBP and TEAM_BASERUN_CS seems to be missing a lot of values which casts doubt on its usefulness as a predictor. Maybe a flag for presense or absense of TEAM_BATTING_HBP and TEAM_BASERUN_CS might be a better predictor. Also given the fact that there is low correlation, we decided to exclude these 2 variables from any missing value or outlier treatment.

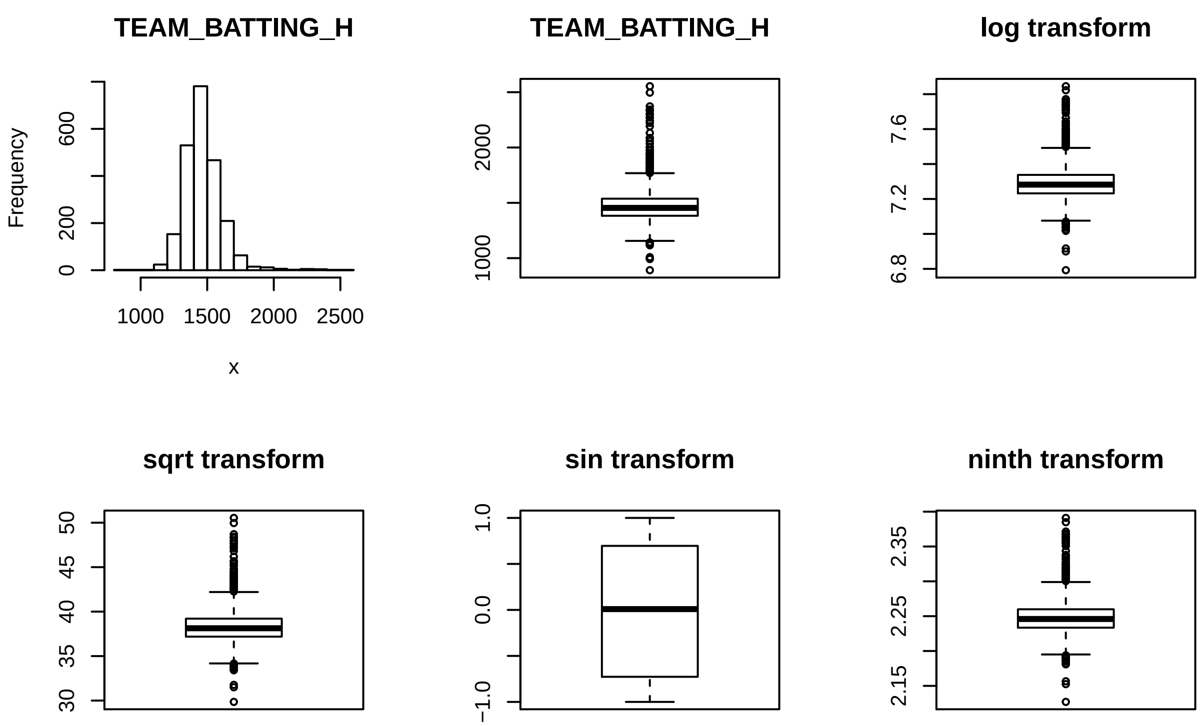
3.Most of the variables seem to indicate a positive / negative correlation in line with the theoretical effect. However, the following stand out as they show a correlation opposite to the theoretical impact: TEAM_BASERUN_CS, TEAM_PITCHING_HR, TEAM_PITCHING_BB, TEAM_PITCHING_SO and TEAM_FIELDING_DP. Lets evaluate these variables further once we fix any missing values or outliers.

4. We will impute the missing values in TEAM_BATTING_SO, FIELDING_DP, BASERUN_SB and TEAM_PITCHING_SO since it has lesser missing values even though there is low correlation. So we will create new variables that will have the respective missing values handled.

1.3 Outliers and Missing Values Identification

In this section we look at boxplots to determine the outliers in variables and decide on whether to act on the outliers.

Lets do some univariate analysis. We will look at the Histogram and Boxplot for each variable to detect outliers if any and treat it accordingly.



For TEAM_BATTING_H, we can see that there are quite a few outliers, both at the upper and lower end. Accordingly, we decide to create a new variable that will have the outlier fixed.

***Please note that we have created similar figures to figure 1 above for each remaining variable. However, we hid the remaining figures for ease of streamlining the report as they have similar shapes. However, we have drawn the below observations from each remaining figure.

For TEAM_BATTING_2B, we can see that there are quite a few outliers, both at the upper and a single outlier at the lower end. For this variable we decide to create a new variable that will have the outliers fixed.

For TEAM_BATTING_3B, we can see that there are quite a few outliers at the upper end. For this variable we decide to create a new variable that will have the outliers fixed.

For TEAM_BATTING_HR, we can see that there are no outliers.

For TEAM_BATTING_BB, we can see that there are quite a few outliers, both at the upper and lower end. For this variable we decide to create a new variable that will have the outlier fixed.

For TEAM_BATTING_SO, we can see that there are no outliers. No further action needed for this variable.

For TEAM_BASERUN_SB, we can see that there are quite a few outliers at the upper end. For this variable we decide to create a new variable that will have the outlier fixed.

For TEAM_FIELDING_E, we can see that there are quite a few outliers at the upper end. For this variable we decide to create a new variable that will have the outlier fixed.

For TEAM_FIELDING_DP, we can see that there are quite a few outliers, both at the upper and lower end. For this variable we decide to create a new variable that will have the outlier fixed.

For TEAM_PITCHING_BB, we can see that there are quite a few outliers, both at the upper and lower end. For this variable we decide to create a new variable that will have the outlier fixed.

For TEAM_PITCHING_H, we can see that there are quite a few outliers at the upper end. For this variable we decide to create a new variable that will have the outlier fixed.

For TEAM_PITCHING_HR, we can see that there only 3 outliers at the upper end. For this variable we decide to create a new variable that will have the outlier fixed.

For TEAM_PITCHING_SO, we can see that there are quite a few outliers at the upper and a single outlier on the lower end. For this variable we decide to create a new variable that will have the outlier fixed.

Please note that, in most of the cases above, we see that a SIN transformation seems to work well to take care of the outliers. We will go ahead and create these new variables respectively.

2. Data Preparation

Now that we have completed the preliminary analysis, we will be cleaning and consolidating data into one dataset for use in analysis and modeling. We will be puring the belwo steps as guidlines:
- Outliers treatment
- Missing values treatment
- Data transformation

2.1 Outliers treatment

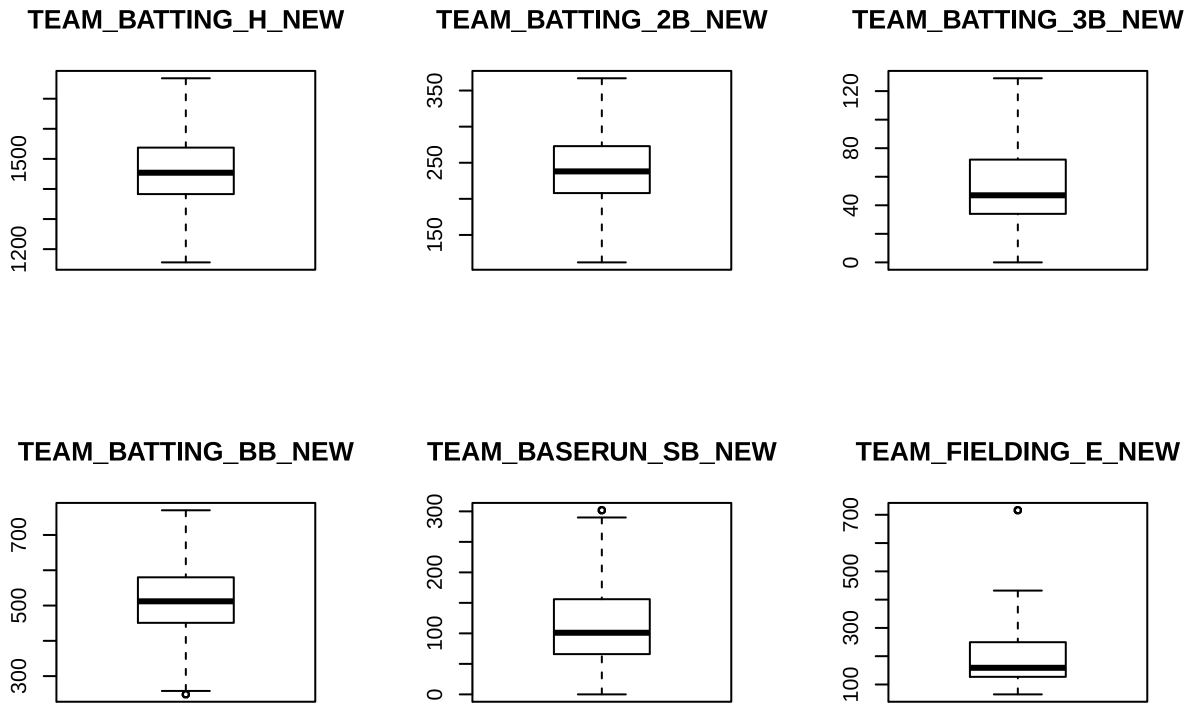
For outliers, we will create 2 sets of variables.

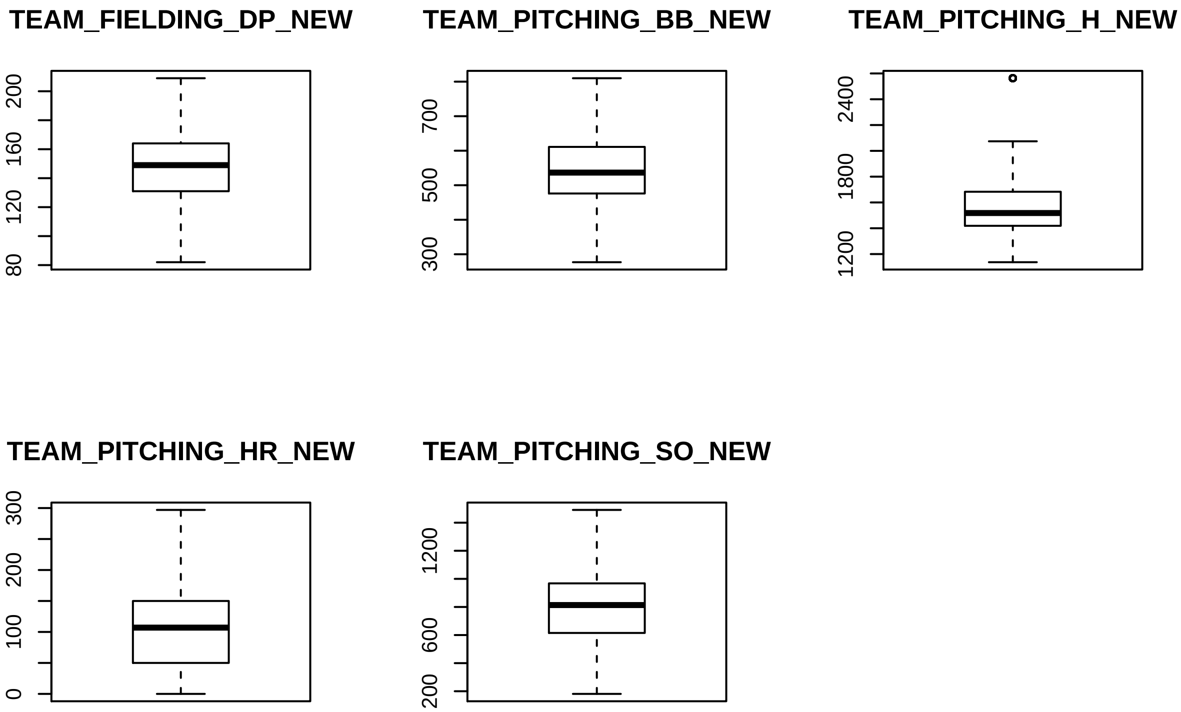
The first set uses the capping method. In this method, we will replace all outliers that lie outside the 1.5 times of IQR limits. We will cap it by replacing those observations less than the lower limit with the value of 5th %ile and those that lie above the upper limit with the value of 95th %ile.

Accordingly we create the following new variables while retaining the original variables.

TEAM_BATTING_H_NEW
TEAM_BATTING_2B_NEW
TEAM_BATTING_3B_NEW
TEAM_BATTING_BB_NEW
TEAM_BASERUN_SB_NEW
TEAM_FIELDING_E_NEW
TEAM_FIELDING_DP_NEW
TEAM_PITCHING_BB_NEW
TEAM_PITCHING_H_NEW
TEAM_PITCHING_HR_NEW
TEAM_PITCHING_SO_NEW

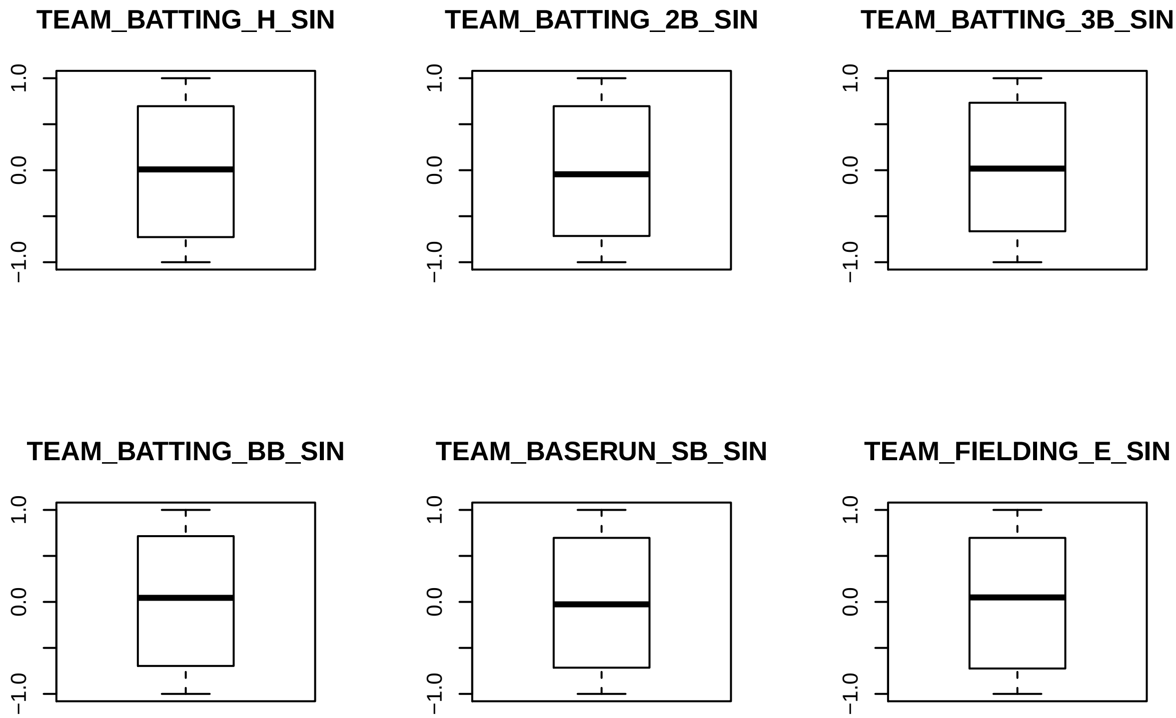
Lets see how the new variables look in boxplots.

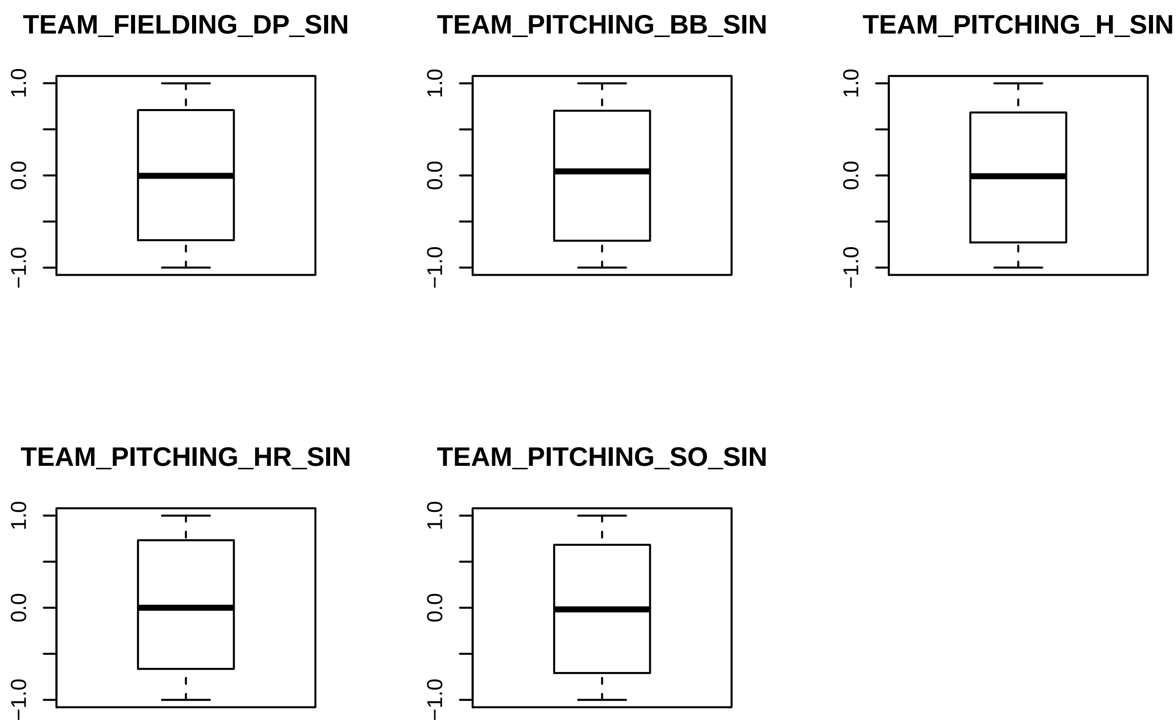

TEAM_BATTING_H_NEW
1200
1500
TEAM_BATTING_2B_NEW
150
250
350
TEAM_BATTING_3B_NEW
0
40
80
120
TEAM_BATTING_BB_NEW
300
500
700
TEAM_BASERUN_SB_NEW
0
100
200
300
TEAM_FIELDING_E_NEW
100
300
500
700



In the second set, we will use the sin transformation and create the following variables:

TEAM_BATTING_H_SIN
TEAM_BATTING_2B_SIN
TEAM_BATTING_3B_SIN
TEAM_BATTING_BB_SIN
TEAM_BASERUN_SB_SIN
TEAM_FIELDING_E_SIN
TEAM_FIELDING_DP_SIN
TEAM_PITCHING_BB_SIN
TEAM_PITCHING_H_SIN
TEAM_PITCHING_HR_SIN
TEAM_PITCHING_SO_SIN


TEAM_BATTING_H_SIN
TEAM_BATTING_2B_SIN
TEAM_BATTING_3B_SIN
TEAM_BATTING_BB_SIN
TEAM_BASERUN_SB_SIN
TEAM_FIELDING_E_SIN
1.0
0.0
−1.0



2.2 Missing values treatment

Next we impute missing values. Since we have handled outliers, we can go ahead and use the mean as impute values. As with outliers, we will go ahead and create new variables for the following:

TEAM_BATTING_SO_NEW

We will re-use the already created new variables for fixing the missing values for the below:

TEAM_PITCHING_SO_NEW
TEAM_BASERUN_SB_NEW
TEAM_FIELDING_DP_NEW

Lets now create some additional variables that might help us in out analysis.

2.3 Missing Flags

First we create flag variables to indicate whether TEAM_BATTING_HBP and TEAM_BASERUN_CS and missing. If the value is missing, we code it with 1 and if the value is present we code it with 0.
We will name our missing flag variables as follow:
TEAM_BATTING_HBP_Missing
TEAM_BASERUN_CS_Missing

2.4 Ratios

Next we create some additional variables, that we think may be useful with the prediction. Here we create the following ratios:
Hits_R = TEAM_BATTING_H/TEAM_PITCHING_H
Walks_R = TEAM_BATTING_BB/TEAM_PITCHING_BB
HomeRuns_R = TEAM_BATTING_HR/TEAM_PITCHING_HR
Strikeout_R = TEAM_BATTING_SO/TEAM_PITCHING_SO

2.5 Calculated Variables

Finally, we will also create calculated variables as below:

1. TEAM_BATTING_EB (Extra Base Hits) = 2B + 3B + HR
2. TEAM_BATTING_1B (Singles by batters) = TEAM_BATTING_H - TEAM_BATTING_EB

2.6 Correlation for new variables

Lets see how the new variables stack up against wins.

Table 4: New variables Correlation

TEAM_BATTING_HBP_Missing	0.0026106
TEAM_BASERUN_CS_Missing	0.0048642
Hits_R	0.0958000
Walks_R	0.0836602
HomeRuns_R	0.0134410
Strikeout_R	0.0631939
TEAM_BATTING_EB	0.3449581
TEAM_BATTING_1B	0.2174301

All new variables seem to have a positive correlation with wins. However, some of them do not seem to have a strong correlation. Lets see how they perform while modeling.

3 Build Models

In this phase, we will build four models. The models independent variables will be based initially on the original data set variables, derived dataset variables, transformed dataset variables, and all variables in the dataset. In addition, for each model, we will perform a stepwise selection and stop at a point where we retain only those variables that have lower AIC (Akaike An Information Criterion). Recall (AIC) is a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Lower AIC leads to better quality model.

Below is a summary table showing models and their respective variables.

VARIABLE_NAME	Comments	Theoretical.Effect	Model1	Model2	Model3	Model4
TEAM_BATTING_H	Given	Positive	Y			Y
TEAM_BATTING_2B	Given	Positive	Y			Y
TEAM_BATTING_3B	Given	Positive	Y			Y
TEAM_BATTING_HR	Given	Positive	Y			Y
TEAM_BATTING_BB	Given	Positive	Y			Y
TEAM_BATTING_HBP	Given	Positive	Y			
TEAM_BATTING_SO	Given	Negative	Y			Y
TEAM_BASERUN_SB	Given	Positive	Y			Y
TEAM_BASERUN_CS	Given	Negative	Y			
TEAM_FIELDING_E	Given	Negative	Y			Y
TEAM_FIELDING_DP	Given	Positive	Y			Y
TEAM_PITCHING_BB	Given	Negative	Y			Y
TEAM_PITCHING_H	Given	Negative	Y			Y
TEAM_PITCHING_HR	Given	Negative	Y			Y
TEAM_PITCHING_SO	Given	Positive	Y			Y
TEAM_BATTING_H_NEW	Derived	Positive		Y		Y
TEAM_BATTING_2B_NEW	Derived	Positive		Y		Y
TEAM_BATTING_3B_NEW	Derived	Positive		Y		Y
TEAM_BATTING_BB_NEW	Derived	Positive		Y		Y
TEAM_BASERUN_SB_NEW	Derived	Positive		Y		Y
TEAM_FIELDING_E_NEW	Derived	Negative		Y		Y
TEAM_FIELDING_DP_NEW	Derived	Positive		Y		Y
TEAM_PITCHING_BB_NEW	Derived	Negative		Y		Y
TEAM_PITCHING_H_NEW	Derived	Negative		Y		Y
TEAM_PITCHING_HR_NEW	Derived	Negative		Y		Y
TEAM_PITCHING_SO_NEW	Derived	Positive		Y		Y
TEAM_BATTING_H_SIN	Derived	Positive			Y	Y
TEAM_BATTING_2B_SIN	Derived	Positive			Y	Y
TEAM_BATTING_3B_SIN	Derived	Positive			Y	Y
TEAM_BATTING_BB_SIN	Derived	Positive			Y	Y
TEAM_BASERUN_SB_SIN	Derived	Positive			Y	Y
TEAM_FIELDING_E_SIN	Derived	Negative			Y	Y
TEAM_FIELDING_DP_SIN	Derived	Positive			Y	Y
TEAM_PITCHING_BB_SIN	Derived	Negative			Y	Y
TEAM_PITCHING_H_SIN	Derived	Negative			Y	Y
TEAM_PITCHING_HR_SIN	Derived	Negative			Y	Y
TEAM_PITCHING_SO_SIN	Derived	Positive			Y	Y
TEAM_BATTING_HBP_Missing	Derived				Y	Y
TEAM_BASERUN_CS_Missing	Derived				Y	Y
Hits_R	Derived				Y	Y
Walks_R	Derived				Y	Y
HomeRuns_R	Derived				Y	Y
Strikeout_R	Derived				Y	Y
TEAM_BATTING_EB	Derived				Y	Y
TEAM_BATTING_1B	Derived				Y	Y

3.1 Model One

In this model, we will be using the original variables. We will create model and we will highlight the variables that being recommended using the AIC value.
First we will produce the summary model as per below:

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_HBP +
##     TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_BASERUN_CS + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP + TEAM_PITCHING_BB + TEAM_PITCHING_H + TEAM_PITCHING_HR +
##     TEAM_PITCHING_SO, data = na.omit(moneyball2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8708  -5.6564  -0.0599   5.2545  22.9274
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      60.28826   19.67842   3.064  0.00253 **
## TEAM_BATTING_H    1.91348    2.76139   0.693  0.48927
## TEAM_BATTING_2B   0.02639    0.03029   0.871  0.38484
## TEAM_BATTING_3B  -0.10118    0.07751  -1.305  0.19348
## TEAM_BATTING_HR  -4.84371   10.50851  -0.461  0.64542
## TEAM_BATTING_BB  -4.45969    3.63624  -1.226  0.22167
## TEAM_BATTING_HBP  0.08247    0.04960   1.663  0.09815 .
## TEAM_BATTING_SO   0.34196    2.59876   0.132  0.89546
## TEAM_BASERUN_SB   0.03304    0.02867   1.152  0.25071
## TEAM_BASERUN_CS  -0.01104    0.07143  -0.155  0.87730
## TEAM_FIELDING_E  -0.17204    0.04140  -4.155 5.08e-05 ***
## TEAM_FIELDING_DP -0.10819    0.03654  -2.961  0.00349 **
## TEAM_PITCHING_BB  4.51089    3.63372   1.241  0.21612
## TEAM_PITCHING_H  -1.89096    2.76095  -0.685  0.49432
## TEAM_PITCHING_HR  4.93043   10.50664   0.469  0.63946
## TEAM_PITCHING_SO -0.37364    2.59705  -0.144  0.88577
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.467 on 175 degrees of freedom
## Multiple R-squared:  0.5501, Adjusted R-squared:  0.5116
## F-statistic: 14.27 on 15 and 175 DF,  p-value: < 2.2e-16
```

We notice that model 1 has the following summary characteristics:
-The Residual standard error is 8.467
-Degrees of freedom: 175
-Deleted observations due missing data: 2085.
-Multiple R-squared: 0.5501
-Adjusted R-squared: 0.5116
-F-statistic: 14.27 on 15 and 175 DF
-p-value: < 2.2e-16

Next. we will step thru this model (model 1) and retain only those variables that have the most impact. below the relevant varuibale for model 1:

	Coefficients
(Intercept)	60.9545372
TEAM_BATTING_H	0.0254136
TEAM_BATTING_HBP	0.0871197
TEAM_FIELDING_E	-0.1721804
TEAM_FIELDING_DP	-0.1190433
TEAM_PITCHING_BB	0.0567223
TEAM_PITCHING_HR	0.0894498
TEAM_PITCHING_SO	-0.0313631

3.2 Model Two

In this model (model2), we will be using the adjusted values based on our outlier treatment process. We will create model and we will highlight the variables that being recommended using the AIC value. First we will produce the summary model as per below:

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H_NEW + TEAM_BATTING_2B_NEW +
##     TEAM_BATTING_3B_NEW + TEAM_BATTING_BB_NEW + TEAM_BASERUN_SB_NEW +
##     TEAM_FIELDING_E_NEW + TEAM_FIELDING_DP_NEW + TEAM_PITCHING_BB_NEW +
##     TEAM_PITCHING_H_NEW + TEAM_PITCHING_HR_NEW + TEAM_PITCHING_SO_NEW,
##     data = na.omit(moneyball2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.8141  -6.3893  -0.0595   5.0336  22.0504
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          58.83398   19.34512   3.041 0.002710 **
## TEAM_BATTING_H_NEW   -0.10194    0.20504  -0.497 0.619664
## TEAM_BATTING_2B_NEW   0.02566    0.03072   0.835 0.404644
## TEAM_BATTING_3B_NEW  -0.12553    0.07569  -1.658 0.098993 .
## TEAM_BATTING_BB_NEW   0.03674    0.08499   0.432 0.666031
## TEAM_BASERUN_SB_NEW   0.03137    0.02271   1.381 0.168873
## TEAM_FIELDING_E_NEW  -0.17714    0.04048  -4.376 2.05e-05 ***
## TEAM_FIELDING_DP_NEW -0.10377    0.03657  -2.838 0.005070 **
## TEAM_PITCHING_BB_NEW  0.01763    0.08317   0.212 0.832365
## TEAM_PITCHING_H_NEW   0.12603    0.20539   0.614 0.540252
## TEAM_PITCHING_HR_NEW  0.09054    0.02564   3.532 0.000525 ***
## TEAM_PITCHING_SO_NEW -0.02961    0.00731  -4.051 7.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.469 on 179 degrees of freedom
## Multiple R-squared:  0.5396, Adjusted R-squared:  0.5113
## F-statistic: 19.07 on 11 and 179 DF,  p-value: < 2.2e-16
```

Lets now step thru this model and retain only those variables that have the most impact.

	Coefficients
(Intercept)	59.6641806
TEAM_BATTING_3B_NEW	-0.1220735
TEAM_BATTING_BB_NEW	0.0550034
TEAM_FIELDING_E_NEW	-0.1742357
TEAM_FIELDING_DP_NEW	-0.1123065
TEAM_PITCHING_H_NEW	0.0313496
TEAM_PITCHING_HR_NEW	0.0809384
TEAM_PITCHING_SO_NEW	-0.0284152

3.3 Model Three

In this model (model3), we will be using the derived values based on our variable transformation process. We will create model and we will highlight the variables that being recommended using the AIC value. First we will produce the summary model as per below:

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H_SIN + TEAM_BATTING_2B_SIN +
##     TEAM_BATTING_3B_SIN + TEAM_BATTING_BB_SIN + TEAM_BASERUN_SB_SIN +
##     TEAM_FIELDING_E_SIN + TEAM_FIELDING_DP_SIN + TEAM_PITCHING_BB_SIN +
##     TEAM_PITCHING_H_SIN + TEAM_PITCHING_HR_SIN + TEAM_PITCHING_SO_SIN +
##     TEAM_BATTING_HBP_Missing + TEAM_BASERUN_CS_Missing + Hits_R +
##     Walks_R + HomeRuns_R + Strikeout_R + TEAM_BATTING_EB + TEAM_BATTING_1B,
##     data = na.omit(moneyball2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.0633  -7.2221   0.1263   6.9949  24.0791
##
## Coefficients: (2 not defined because of singularities)
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.557e+02  4.035e+02   1.129  0.26031
## TEAM_BATTING_H_SIN       -6.147e-01  2.083e+00  -0.295  0.76823
## TEAM_BATTING_2B_SIN       8.235e-02  1.085e+00   0.076  0.93960
## TEAM_BATTING_3B_SIN       5.884e-01  1.144e+00   0.514  0.60772
## TEAM_BATTING_BB_SIN      -2.319e+00  2.291e+00  -1.012  0.31295
## TEAM_BASERUN_SB_SIN      -2.182e+00  1.104e+00  -1.978  0.04957 *
## TEAM_FIELDING_E_SIN       5.054e-01  1.099e+00   0.460  0.64625
## TEAM_FIELDING_DP_SIN      2.355e+00  1.115e+00   2.113  0.03602 *
## TEAM_PITCHING_BB_SIN      4.716e-01  2.246e+00   0.210  0.83392
## TEAM_PITCHING_H_SIN       7.726e-01  2.068e+00   0.374  0.70920
## TEAM_PITCHING_HR_SIN     -1.696e+00  1.101e+00  -1.541  0.12526
## TEAM_PITCHING_SO_SIN      7.777e-01  1.113e+00   0.699  0.48564
## TEAM_BATTING_HBP_Missing         NA         NA      NA       NA
## TEAM_BASERUN_CS_Missing          NA         NA      NA       NA
## Hits_R                    4.799e+02  9.617e+03   0.050  0.96026
## Walks_R                  -1.007e+04  5.217e+03  -1.930  0.05524 .
## HomeRuns_R                3.948e+03  2.006e+03   1.968  0.05068 .
```

```
## Strikeout_R              5.172e+03  8.565e+03   0.604  0.54673
## TEAM_BATTING_EB          1.020e-01  1.756e-02   5.812  2.9e-08 ***
## TEAM_BATTING_1B          4.287e-02  1.298e-02   3.303  0.00116 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.14 on 173 degrees of freedom
## Multiple R-squared:  0.3622, Adjusted R-squared:  0.2995
## F-statistic: 5.778 on 17 and 173 DF,  p-value: 2.536e-10
```

Lets now step thru this model and retain only those variables that have the most impact.

	Coefficients
(Intercept)	407.3259680
TEAM_BATTING_BB_SIN	-2.0075801
TEAM_BASERUN_SB_SIN	-2.2016077
TEAM_FIELDING_DP_SIN	2.3971038
TEAM_PITCHING_HR_SIN	-1.7781816
Walks_R	-5393.0576151
HomeRuns_R	4971.9677657
TEAM_BATTING_EB	0.1018484
TEAM_BATTING_1B	0.0441886

3.4 Model Four

In this model (model4), we will be using all variables original, adjusted, and derived values. We will create model and we will highlight the variables that being recommended using the AIC value. First we will produce the summary model as per below:

```
##
## Coefficients: (14 not defined because of singularities)
##                         Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)           1.5803e+04  1.6334e+04  0.9675   0.33478
## TEAM_BATTING_H        1.4193e+01  1.1598e+01  1.2238   0.22286
## TEAM_BATTING_2B       6.9092e-02  1.2259e-01  0.5636   0.57382
## TEAM_BATTING_3B      -6.7019e-02  8.0344e-02 -0.8341   0.40546
## TEAM_BATTING_HR      -3.1382e+01  2.2076e+01 -1.4216   0.15712
## TEAM_BATTING_BB       1.4865e+01  8.5333e+00  1.7420   0.08345
## TEAM_BATTING_SO      -7.5314e+00  3.9562e+00 -1.9037   0.05877
## TEAM_BASERUN_SB       2.8975e-02  3.0059e-02  0.9639   0.33655
## TEAM_BASERUN_CS      -2.5752e-02  7.2907e-02 -0.3532   0.72440
## TEAM_BATTING_HBP      8.8491e-02  5.0536e-02  1.7510   0.08188
## TEAM_PITCHING_H      -1.4180e+01  1.1597e+01 -1.2227   0.22327
## TEAM_PITCHING_HR      3.1466e+01  2.2073e+01  1.4255   0.15597
## TEAM_PITCHING_BB     -1.4843e+01  8.5379e+00 -1.7385   0.08408
## TEAM_PITCHING_SO      7.4970e+00  3.9540e+00  1.8961   0.05978
## TEAM_FIELDING_E      -1.8995e-01  4.3293e-02 -4.3875 2.087e-05
## TEAM_FIELDING_DP     -9.8295e-02  3.8185e-02 -2.5742   0.01097
## TEAM_BATTING_2B_NEW  -4.6960e-02  1.2464e-01 -0.3768   0.70686
```

```
## TEAM_BATTING_BB_NEW   3.1093e-02  8.6932e-02  0.3577   0.72107
## TEAM_BATTING_H_SIN   -8.1802e-01  1.8861e+00 -0.4337   0.66508
## TEAM_BATTING_2B_SIN  -6.8111e-01  9.2777e-01 -0.7341   0.46395
## TEAM_BATTING_3B_SIN  -4.1022e-01  9.8554e-01 -0.4162   0.67780
## TEAM_BATTING_BB_SIN  -1.0084e+00  1.9831e+00 -0.5085   0.61182
## TEAM_BASERUN_SB_SIN  -2.3013e+00  9.3403e-01 -2.4638   0.01482
## TEAM_FIELDING_E_SIN  -4.9238e-01  9.2782e-01 -0.5307   0.59639
## TEAM_FIELDING_DP_SIN  1.7662e+00  9.5433e-01  1.8507   0.06608
## TEAM_PITCHING_BB_SIN -7.4780e-02  1.9432e+00 -0.0385   0.96935
## TEAM_PITCHING_H_SIN   1.0784e+00  1.8692e+00  0.5770   0.56479
## TEAM_PITCHING_HR_SIN -9.5148e-01  9.3622e-01 -1.0163   0.31104
## TEAM_PITCHING_SO_SIN -9.0822e-01  9.6051e-01 -0.9456   0.34581
## Hits_R               -2.3615e+04  2.0532e+04 -1.1502   0.25181
## Walks_R              -1.8042e+04  9.1272e+03 -1.9767   0.04981
## HomeRuns_R            1.1879e+04  6.1102e+03  1.9441   0.05367
## Strikeout_R           1.4052e+04  8.9098e+03  1.5772   0.11675
##
## n = 191, p = 33, Residual SE = 8.27202, R-Squared = 0.61
```

Lets now step thru this model and retain only those variables that have the most impact.

	Coefficients
(Intercept)	-1.818043e+03
TEAM_BATTING_H	1.733820e-02
TEAM_BATTING_BB	8.603592e+00
TEAM_BATTING_SO	-6.535137e+00
TEAM_BATTING_HBP	9.492190e-02
TEAM_PITCHING_HR	8.312810e-02
TEAM_PITCHING_BB	-8.550658e+00
TEAM_PITCHING_SO	6.500379e+00
TEAM_FIELDING_E	-1.810601e-01
TEAM_FIELDING_DP	-1.069442e-01
TEAM_BATTING_BB_SIN	-1.473778e+00
TEAM_BASERUN_SB_SIN	-2.441250e+00
TEAM_FIELDING_DP_SIN	1.896564e+00
Walks_R	-1.474774e+04
HomeRuns_R	4.845243e+03
Strikeout_R	1.179872e+04

Discuss the coefficients in the models, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

4 Model Selection

In section we will further examine all four models. We will apply a model selection strategy by comparing models' AIC, R-squared, and VIF (variance inflation factors. In addition, we will perform diagnostics to validate the assumption of Linear Regression

4.1 Model selection strategy:

Following model selection strategy has been used for this assignment:

(1) Akaike information criterion (AIC) measure has been used to compare relative performance of different models
(2) Along with that of adjusted R^2 values are also used to compare different models performance
(3) Different regression model diagnostics plots has been used to test assumptions for regression- (a) test for normality of residuals (b) plot for randomness of residuals, (c) evaluation of homoscedasticity
(4) Finally model has been tested for collinearity and enhanced by removing collinearity with the use of variance inflation factors (VIF)

Compare models by AIC measures and adjusted R^2 values

```
## [1] 1365.858
```

```
## [1] 1366.497
```

```
## [1] 1430.763
```

```
## [1] 1356.061
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_BB +
##     TEAM_BATTING_SO + TEAM_BATTING_HBP + TEAM_PITCHING_HR + TEAM_PITCHING_BB +
##     TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_BATTING_BB_SIN +
##     TEAM_BASERUN_SB_SIN + TEAM_FIELDING_DP_SIN + Walks_R + HomeRuns_R +
##     Strikeout_R, data = na.omit(moneyball2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.1987  -4.7433   0.0706   4.9994  23.6442
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.818e+03  3.830e+03  -0.475 0.635598
## TEAM_BATTING_H        1.734e-02  1.004e-02   1.726 0.086065 .
## TEAM_BATTING_BB       8.604e+00  6.217e+00   1.384 0.168186
## TEAM_BATTING_SO      -6.535e+00  3.425e+00  -1.908 0.058038 .
## TEAM_BATTING_HBP      9.492e-02  4.706e-02   2.017 0.045220 *
## TEAM_PITCHING_HR      8.313e-02  2.383e-02   3.488 0.000615 ***
## TEAM_PITCHING_BB     -8.551e+00  6.215e+00  -1.376 0.170652
## TEAM_PITCHING_SO      6.500e+00  3.423e+00   1.899 0.059195 .
## TEAM_FIELDING_E      -1.811e-01  3.828e-02  -4.730 4.61e-06 ***
## TEAM_FIELDING_DP     -1.069e-01  3.452e-02  -3.098 0.002267 **
## TEAM_BATTING_BB_SIN  -1.474e+00  8.541e-01  -1.725 0.086212 .
## TEAM_BASERUN_SB_SIN  -2.441e+00  8.511e-01  -2.868 0.004634 **
## TEAM_FIELDING_DP_SIN  1.897e+00  8.531e-01   2.223 0.027491 *
## Walks_R              -1.475e+04  6.742e+03  -2.187 0.030046 *
## HomeRuns_R            4.845e+03  2.381e+03   2.035 0.043350 *
```

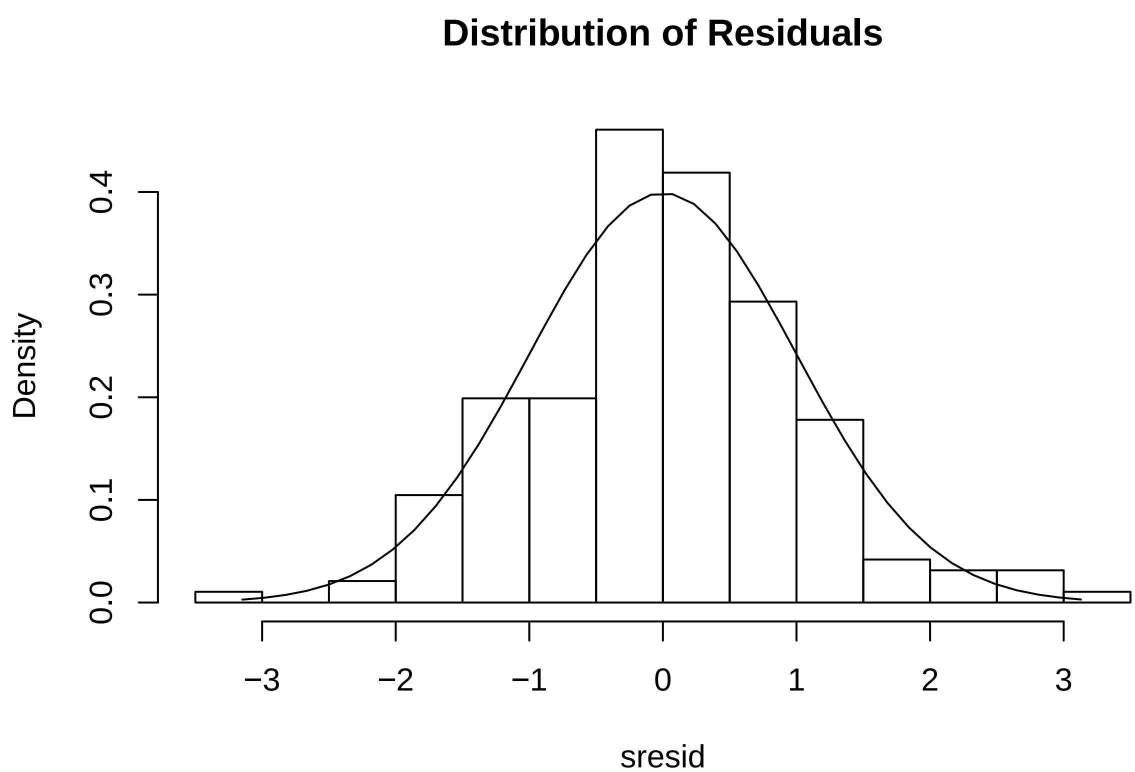
```
## Strikeout_R           1.180e+04  5.360e+03   2.201 0.029020 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.05 on 175 degrees of freedom
## Multiple R-squared:  0.5933, Adjusted R-squared:  0.5585
## F-statistic: 17.02 on 15 and 175 DF,  p-value: < 2.2e-16
```

Looking at the AIC values it appears that models, “step1” & “step 4” are comparatively better models of the pack. “step1” has adjusted R^2 value .5167 which means this model can explain 51.67% variability in data. “step4” has adjusted R^2 value of .5585 and this model can explain 55.85% variability in data. From this two data points model “step4” was picked for further evaluation.

4.2 Model diagnostics

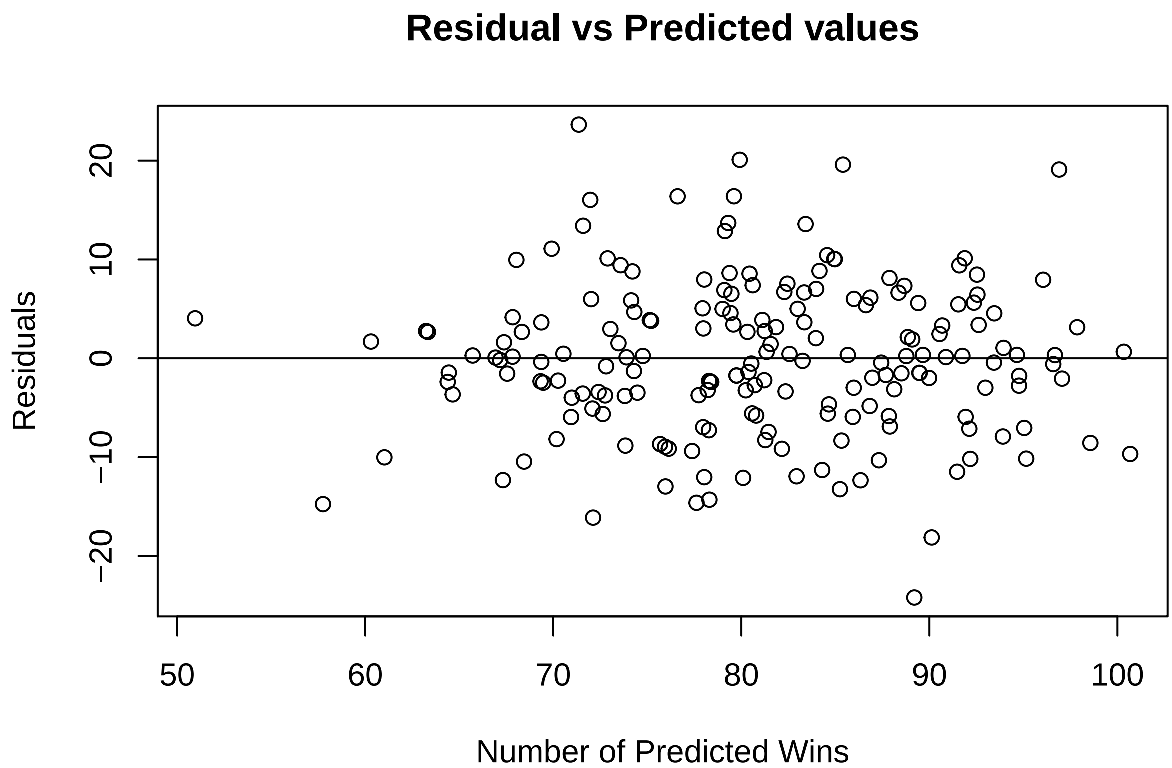
We will create plots to validate the assumption of Linear Regression:

Normality check of residual values:



Based on the normality plot it appears that residual distribution is normal. This indicates the mean of the difference between our predictions.

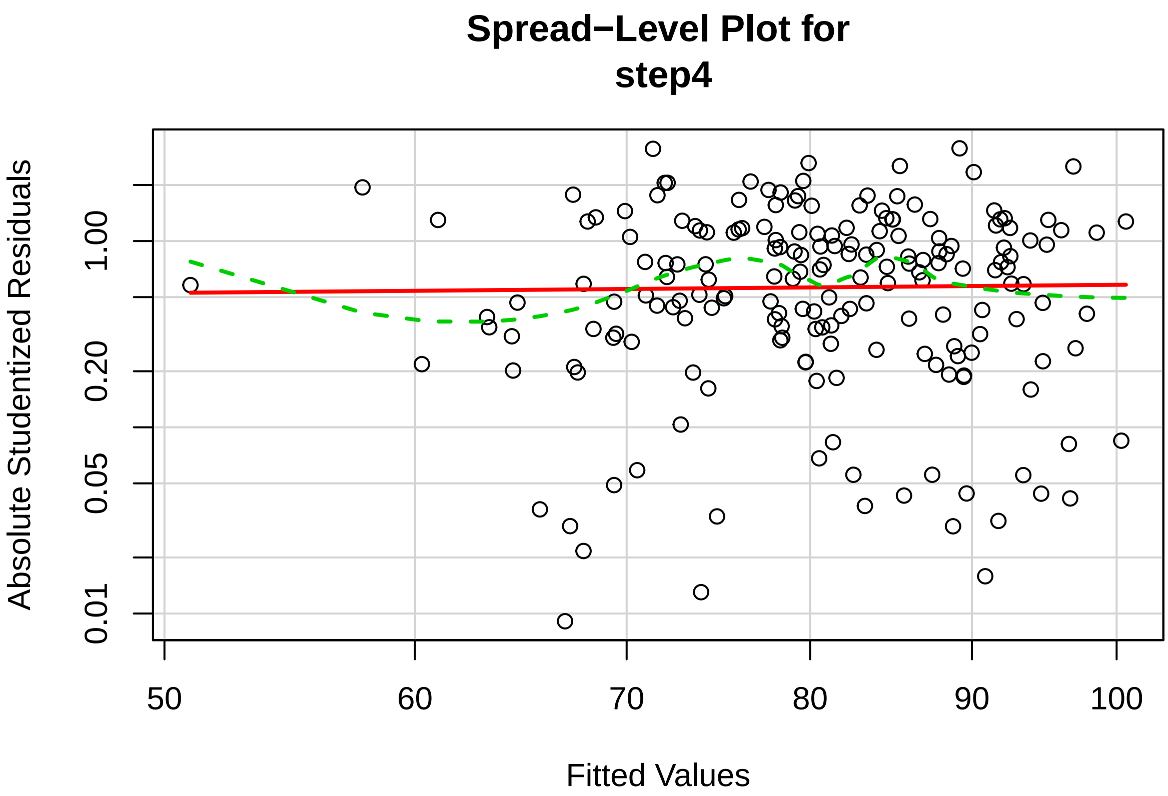
plot residuals with respect to predicted value for randomness:



Distribution of residual values are random around base line and do not show any pattern around base line.

Evaluate homoscedasticity:

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.0009383232    Df = 1     p = 0.975563
```




```
##
## Suggested power transformation:  0.8554837
```

The test confirms the non-constant error variance test. It also has a p-value higher than a significance level of 0.05.

Analysis of collinearity:

Table 10: Analysis of collinearity

TEAM_BATTING_H	1.309556
TEAM_BATTING_BB	796.758220
TEAM_BATTING_SO	610.869636
TEAM_BATTING_HBP	1.044879
TEAM_PITCHING_HR	1.321730
TEAM_PITCHING_BB	797.265244
TEAM_PITCHING_SO	611.561094
TEAM_FIELDING_E	1.090145
TEAM_FIELDING_DP	1.040855
TEAM_BATTING_BB_SIN	1.034981
TEAM_BASERUN_SB_SIN	1.041575
TEAM_FIELDING_DP_SIN	1.041682
Walks_R	24.085235
HomeRuns_R	8.951644

Strikeout_R	19.248743

Variables have been tested with variance inflation factors (VIF). If any variable has value which is greater than 2 then the highest value variable been removed from model and model performance has been evaluated. Following are the out comes from this assessment steps-

pass 1- Based on that variance inflation factors (VIF) following variable "TEAM_PITCHING_BB" has highest value > 2 and is removed from model, and model is evaluated without that variable. Adjusted R^2 value changed from .5585 to .5562. Hence this variable is not adding lot of value to the model and can be removed.

pass 2- Based on that variance inflation factors (VIF) following variable "TEAM_BATTING_SO" has highest value > 2 and is removed from model, and model is evaluated without that variable. Adjusted R^2 values changed from .5562 to .5534. Hence this variable is not adding lot of value to the model and can be removed.

pass 3- Based on that variance inflation factors (VIF) following variable "Strikeout_R" has highest value > 2 and is removed from model, and model is evaluated without that variable. Adjusted R^2 value changed from .5534 to .5526. Hence this variable is not adding lot of value to the model and can be removed.

pass 4- Based on that variance inflation factors (VIF) following variable "HomeRuns_R" has highest value > 2 and is removed from model, and model is evaluated without that variable. Adjusted R^2 value changed from .5526 to .5462 which is some compromise with the performance at the cost of reducing complexity. Reduction of variable will simplify the model and hence updated model is selected with some compromise in performance.

```
##       TEAM_BATTING_H      TEAM_BATTING_BB     TEAM_BATTING_HBP
##             1.269620             1.165108             1.037474
##      TEAM_PITCHING_HR     TEAM_PITCHING_SO      TEAM_FIELDING_E
##             1.294692             1.252542             1.086548
##      TEAM_FIELDING_DP  TEAM_BATTING_BB_SIN  TEAM_BASERUN_SB_SIN
##             1.026350             1.034345             1.039926
## TEAM_FIELDING_DP_SIN              Walks_R
##             1.029959             1.024758
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_BB +
##     TEAM_BATTING_HBP + TEAM_PITCHING_HR + TEAM_PITCHING_SO +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_BATTING_BB_SIN +
##     TEAM_BASERUN_SB_SIN + TEAM_FIELDING_DP_SIN + Walks_R, data = moneyball2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.1241  -5.0179  -0.3098   4.7776  22.5184
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.354e+02  2.899e+02   1.157 0.248862
## TEAM_BATTING_H        2.396e-02  9.871e-03   2.427 0.016215 *
## TEAM_BATTING_BB       5.349e-02  9.217e-03   5.804 2.89e-08 ***
## TEAM_BATTING_HBP      8.191e-02  4.737e-02   1.729 0.085490 .
## TEAM_PITCHING_HR      8.511e-02  2.366e-02   3.596 0.000417 ***
## TEAM_PITCHING_SO     -3.086e-02  7.107e-03  -4.342 2.36e-05 ***
## TEAM_FIELDING_E      -1.810e-01  3.868e-02  -4.679 5.66e-06 ***
## TEAM_FIELDING_DP     -1.168e-01  3.450e-02  -3.385 0.000874 ***
```

```
## TEAM_BATTING_BB_SIN  -1.489e+00  8.654e-01  -1.720 0.087109 .
## TEAM_BASERUN_SB_SIN  -2.434e+00  8.614e-01  -2.826 0.005252 **
## TEAM_FIELDING_DP_SIN  1.578e+00  8.551e-01   1.846 0.066596 .
## Walks_R             -2.695e+02  2.908e+02  -0.927 0.355256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.161 on 179 degrees of freedom
##   (2085 observations deleted due to missingness)
## Multiple R-squared:  0.5725, Adjusted R-squared:  0.5462
## F-statistic: 21.79 on 11 and 179 DF,  p-value: < 2.2e-16
```

Final model was derived after four passes of elimination were carried out. Looking at the VIF values in the final model there is no collinearity among variables < 2. In this scenario a model with slightly less performance was selected to avoid collinearity effect among variables. But there was no significant reduction of model performance. But this exercise helped to reduce the model complexity.

End model Selection