Name: Sachid Deshmukh, Michael Yampol, Vishal Arora, Ann Ferrara

Date: 09-25-2019

Assignment: Data 621 HW-1

Contents

[Data Exploration 1](#_Toc18762933)

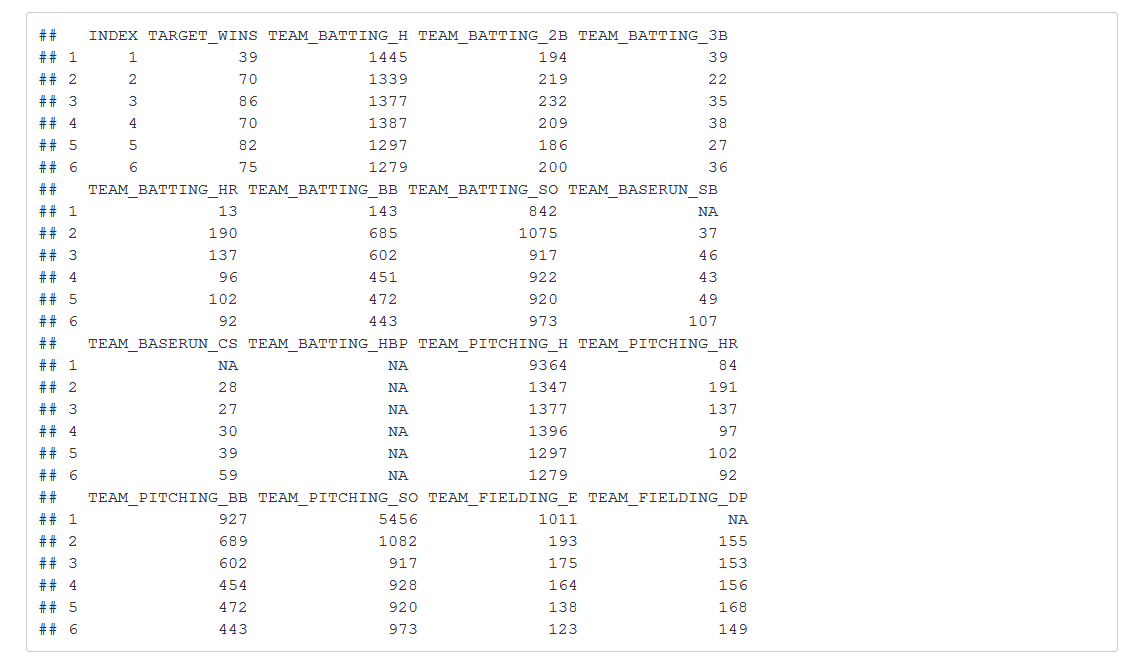
[Data Preparation 4](#_Toc18762934)

[Build Models 6](#_Toc18762935)

[Select Models 13](#_Toc18762936)

# Data Exploration

Let’s load the training dataset and take a preview



Train Data has 2276 observations and 17 variables.

Following columns have missing values.

TEAM\_BATTING\_SO : Strikeouts by batters

TEAM\_BASERUN\_SB : Stolen bases

TEAM\_BASERUN\_CS : Caught stealing

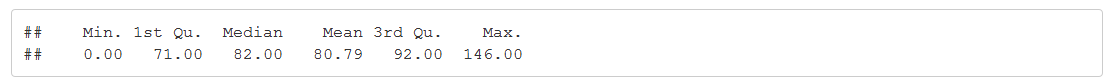
TEAM\_BATTING\_HBP : Batters hit by pitch (get a free base)

TEAM\_PITCHING\_SO : Strikeouts by batters

TEAM\_FIELDING\_DP : Double Plays

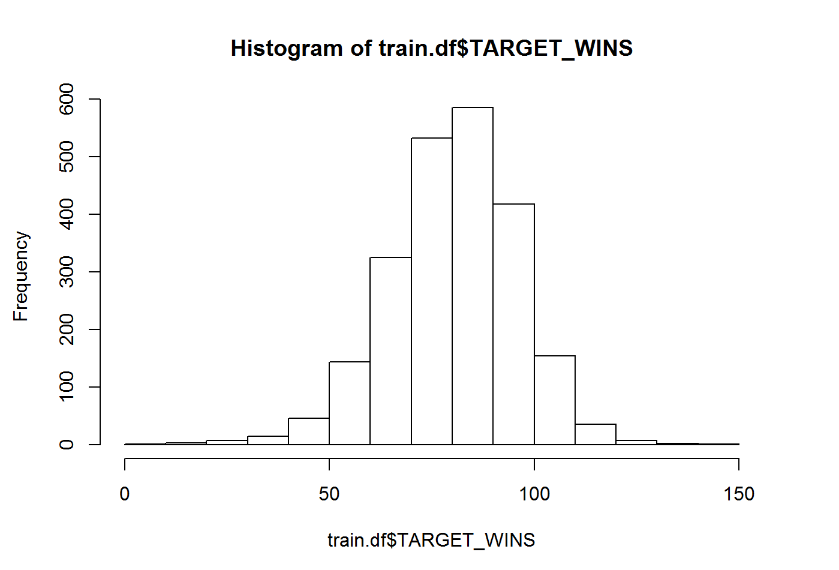
Let’s look at the target variable which is TARGET\_WINS

1] Summary



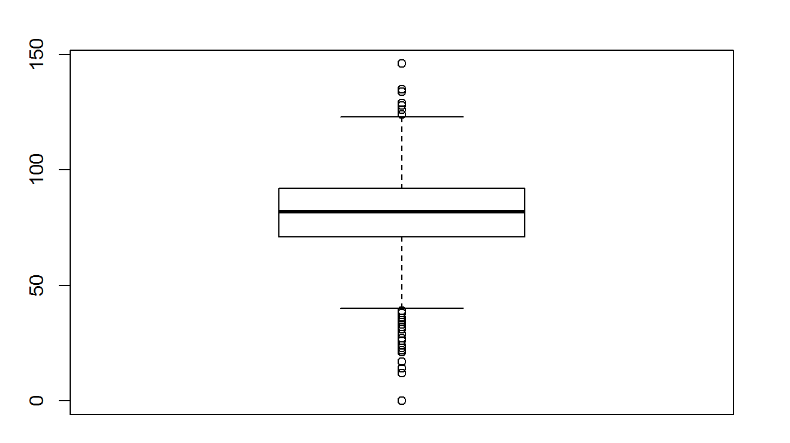
As we can see here Target\_Wins has a range from 0 to 146. Median and Mean values are close to each other indicating there is no skew in the data

2] Distribution

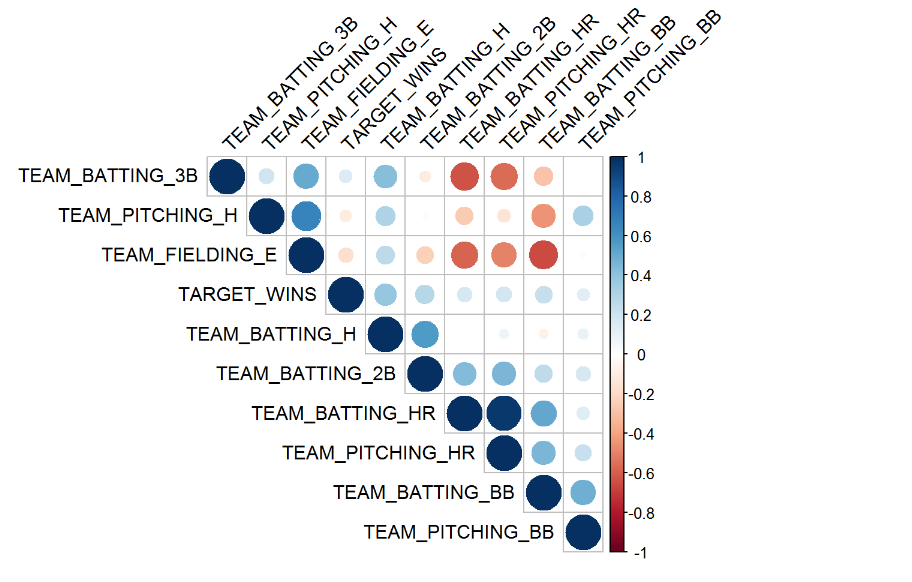


Target variable is normally distributed with mean as 80.79. Histogram above shows that new skew in the data.

3] Box Plot



Box plot for Target\_Wins variable is very symmetric. This proves our point of no data skew. There seems to be two outliers in the box plot. We will keep then in data for now

Let’s look at correlation plot of train dataframe 

Above correlation plot shows that Target\_Wins is highly correlated with following variables

Positive correlation

1] Team\_Batting\_H

2] Team\_Batting\_2B

3] Team\_Batting\_BB

4] Team\_Batting\_HR

5] Team\_Pitching\_HR

6]Team\_Pitching\_BB

Negative Correlation

1] Team\_Fielding\_E

2] Team\_Pitching\_H

# Data Preparation

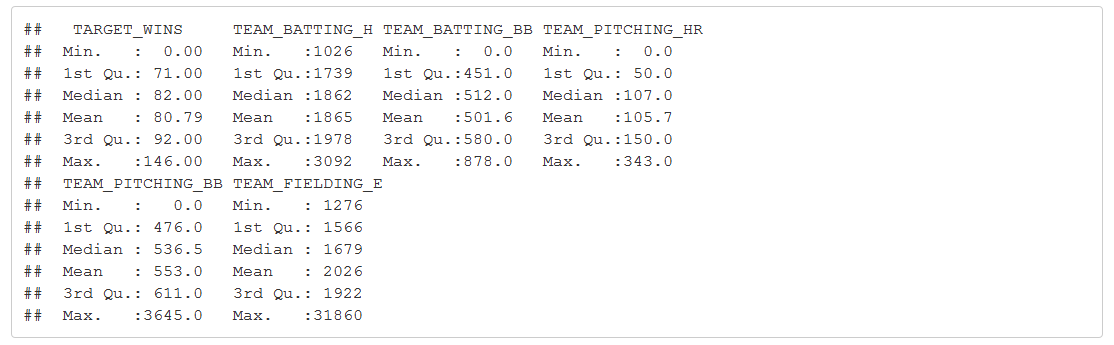
We will do following data clean up and data transformations task on the raw data

1] Drop columns with null values from training data frame

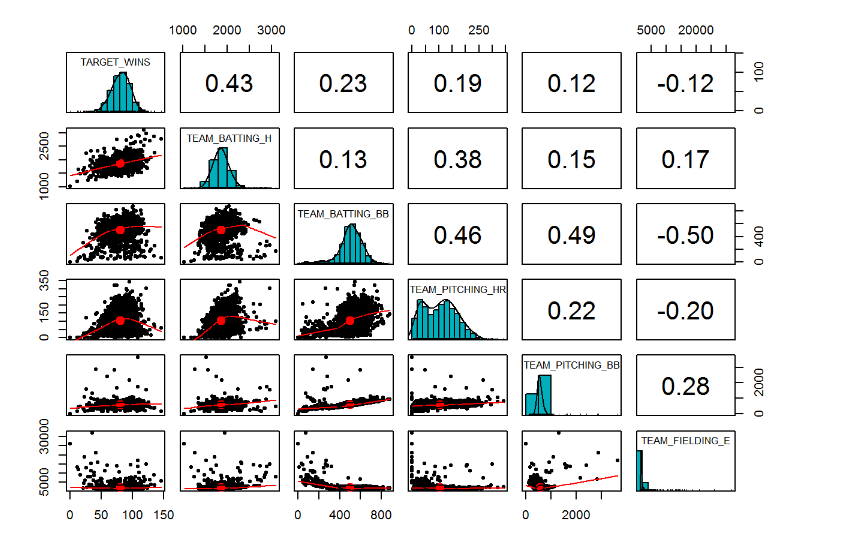
2] Combine Team\_Batting\_H, Team\_Batting\_2B and Team\_Batting\_3B variables into one by taking sum

3] Combine Team\_Fielding\_E and Team\_Pitching\_E variables into one by taking sum

Look at summary of resultant data frame



Look at pair plot (fig 1.0) of resultant data frame



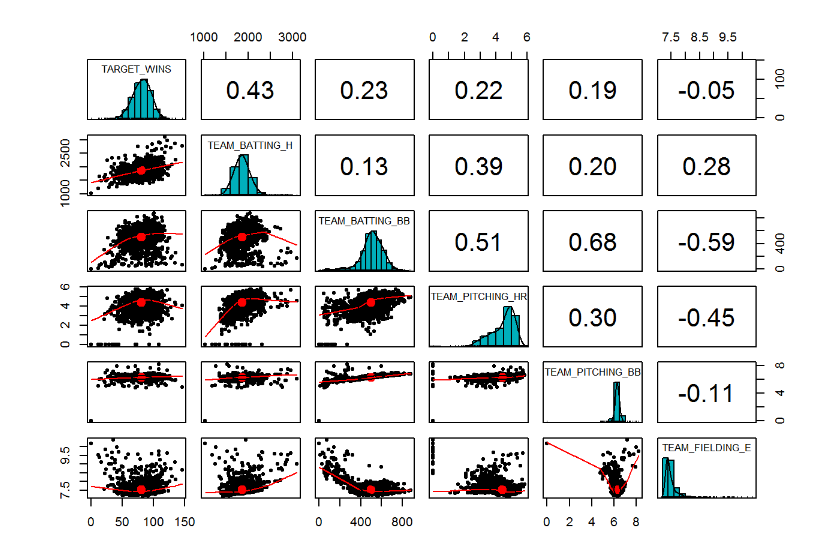
From the pair plot (fig 1.0) above we can see that following variables are not normally distributed.

1] Team\_Pitching\_HR

2] Team\_Pitching\_BB

3] Team\_Fielding\_E

Linear regression model works better if the input variables are normally distributed. Let’s take log transformation of those variables and plot the pair plot (fig 1.0) after transformation



**Figure 1.0**

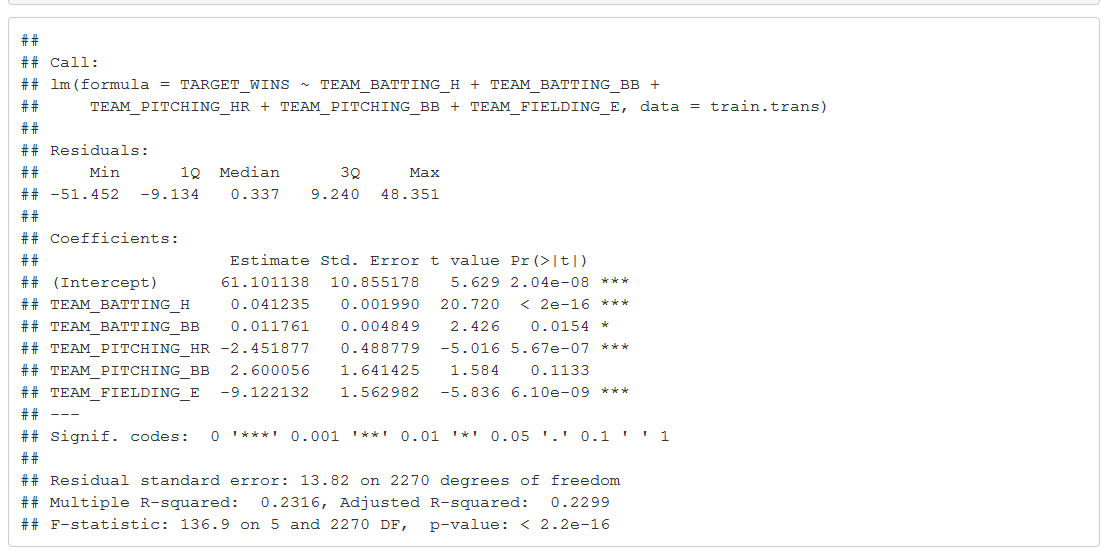
Pair plot after variable transformation looks much better. We can see that all the variables have approximately normal distribution

# Build Models

1] Model-1 In the first model we will model Target\_Wins against following variables

* Team\_Batting\_H
* Team\_Batting\_BB
* Team\_Pitching\_HR
* Team\_Pitching\_BB
* Team\_Fielding\_E

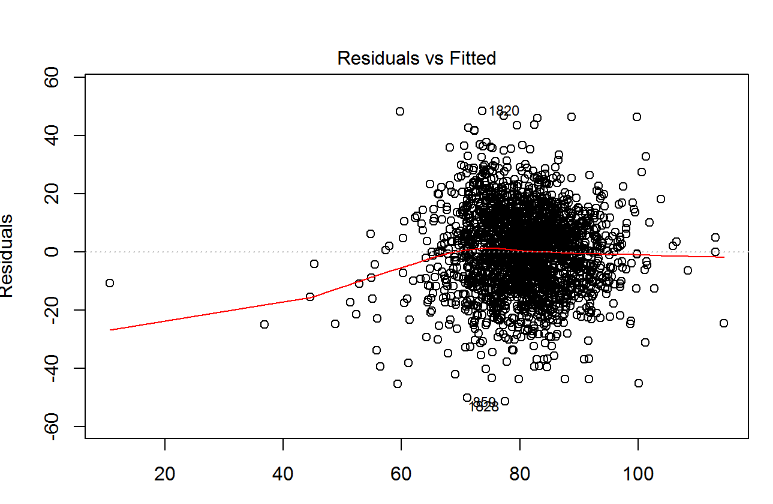
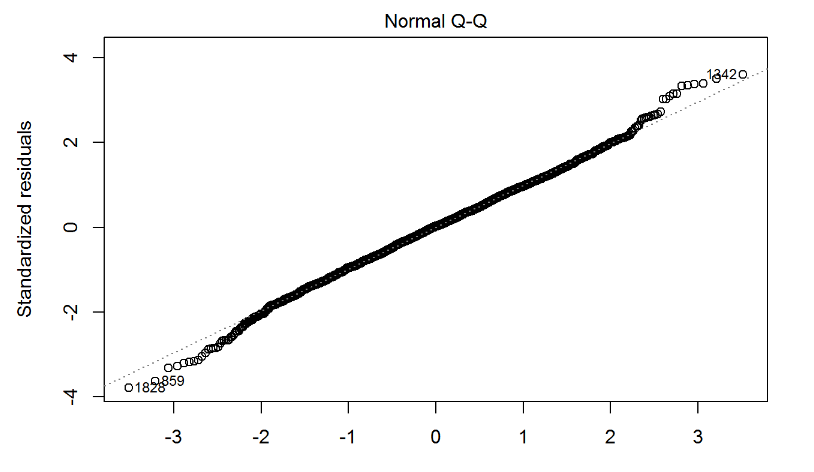
2] Print the model summary



Please note that all the variable coefficients are statistically significant except Team\_PitchingBB. F statistic is significant indicating one or more variables are useful in predicting Target\_Wins variable.

R square value is 0.23 which indicates the model effectiveness

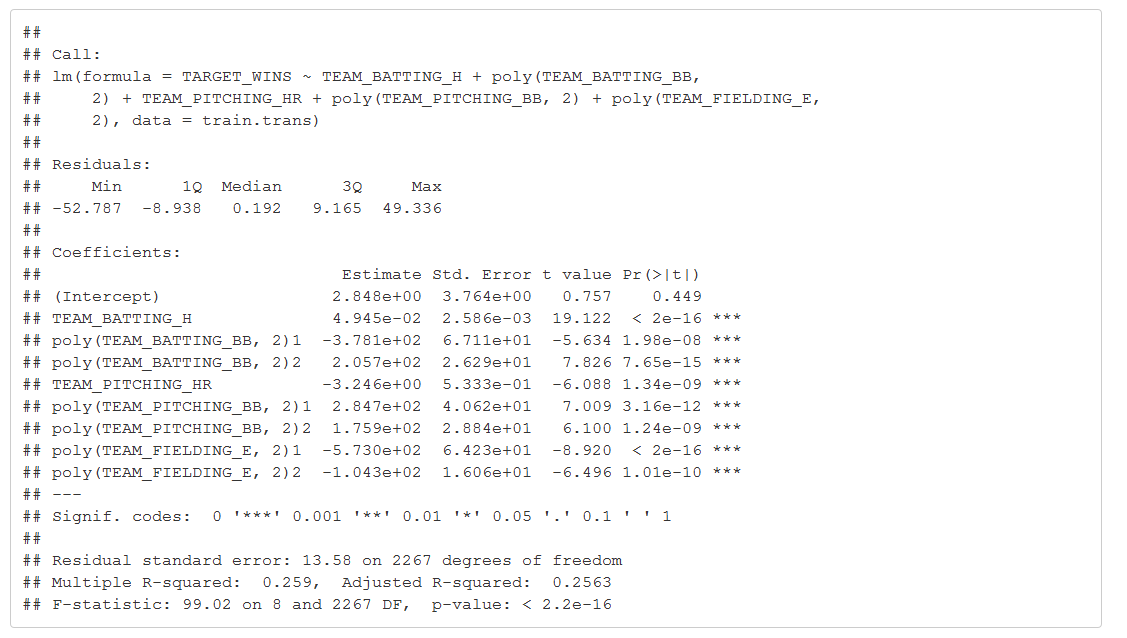
3] Plot the residuals

From the above residual plot we can see that there is some non-linearity between independent variable and dependent variable. We can certainly enhance this model by adding polinimial tearms

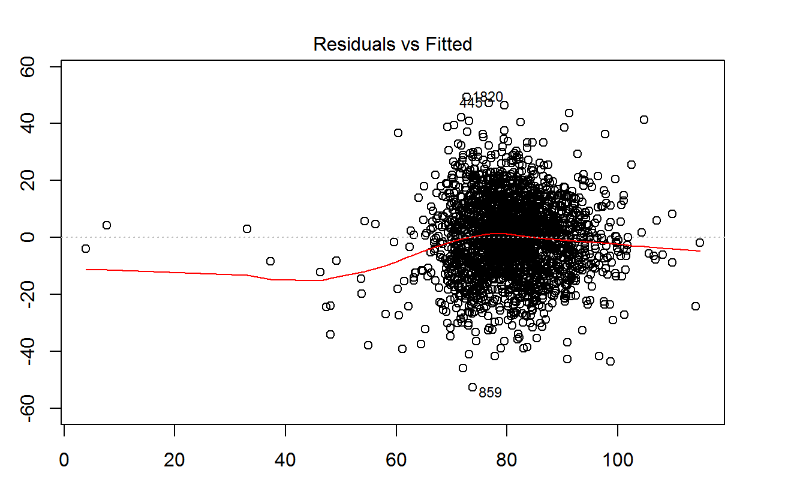
4] Model -2 In the second model from pair plot above (fig 1.0) we can see that there is a non linear relationship between Target\_Wins and Team\_Batting\_BB variable. Let’s add polynomial term for Team\_Batting\_BB variable in the model

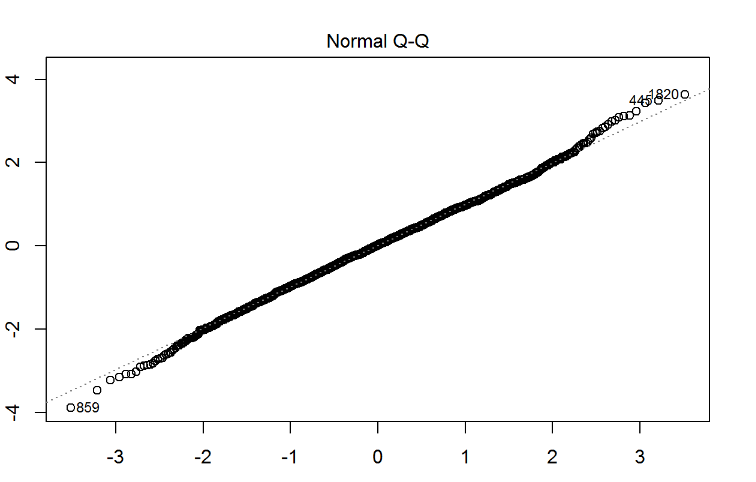
5] Print model summary



We can see from the above model summary that all the coefficient are statistically significant. Adding polinomial term to our model has increased the R square value to 0.259 which is improvement over model-1

6] Plot the residuals



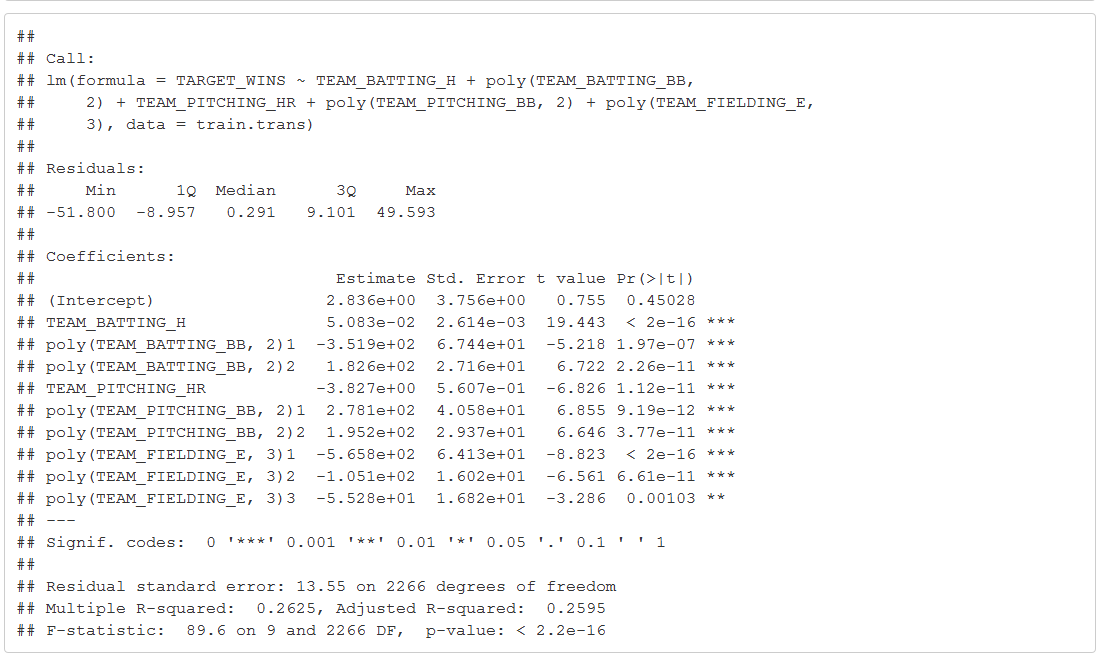


We can see that residuals are looking better now.

7] Model-3 Pair plot above (fig 1.0) shows one more variable which has non-linear relationship with Target\_Wins. Variable Team\_Fielding\_E. This is important variable which is negatively correlated with Target\_Wins. Adding polynomial term for this variable may further improve the model effectiveness

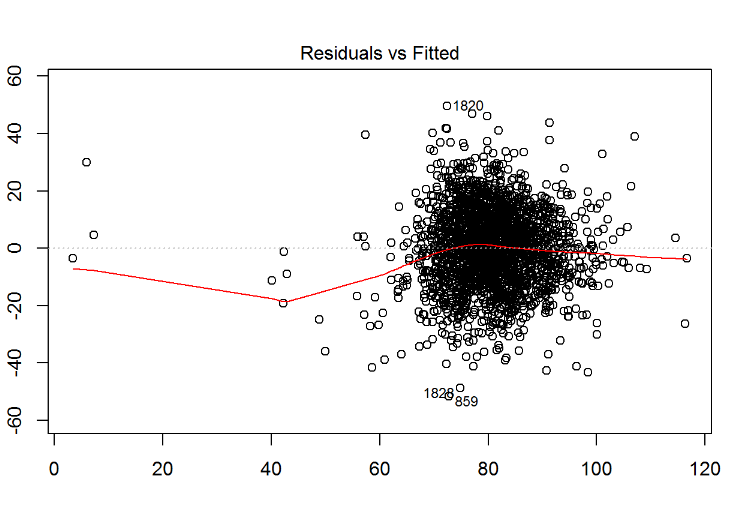
Let’s add polynomial term for Team\_Fielding\_E variable

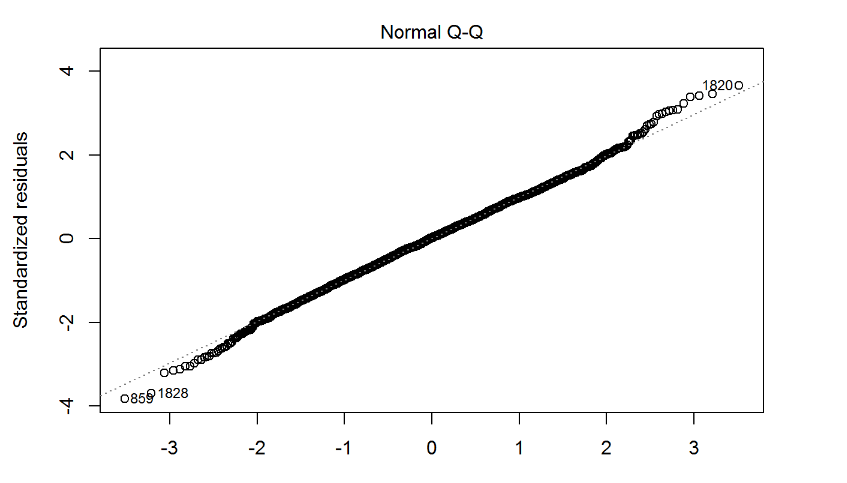
8] Print model summary



From the above model summary we can see that all the coefficient are statistically significant. This indicates that all the variables in the model are effective in predicting Target\_Wins. Statistically significant F stats also proves this point. If we look at the R square it is further improved at 0.2625

9] Print Residual plot





Above residual plot for model-3 looks much better now

# Select Models

For model selection we will use R square as model selection criteria. R square indicates the portion of variance explained by model from total available variance in the data set. Value of R square ranges from 0 to 1. 0 indicates poor model fitting and 1 indicates good fit of the model.

In general model with high R square value indicates better model fit and can be used for model selection.

In the three models above we can see that model-3 has the highest R square 0.2625 so we will select model-3 as our winning model