DATA 621 - Homework 1

Moneyball - Multiple Linear Regression Data Model

Jeyaraman Ramalingam

03/07/2021

Table of Contents

Overview

This document explains the different aspects of the multiple linear regression model built for predicting number of wins for a baseball team with the money ball training dataset provided. The key areas of interest are below. 1. Data Exploration 2. Data Preparation 3. Model Building 4. Model Performance Comparison

1. Data Exploration

The data set given for model building is the money ball training and evaluation datasets.

1. moneyball-training-data.csv
2. moneyball-evaluation-data.csv

To Understand the training dataset , lets use some of the R functions to take a quick peek at the input dataset

Data First look and brief summary

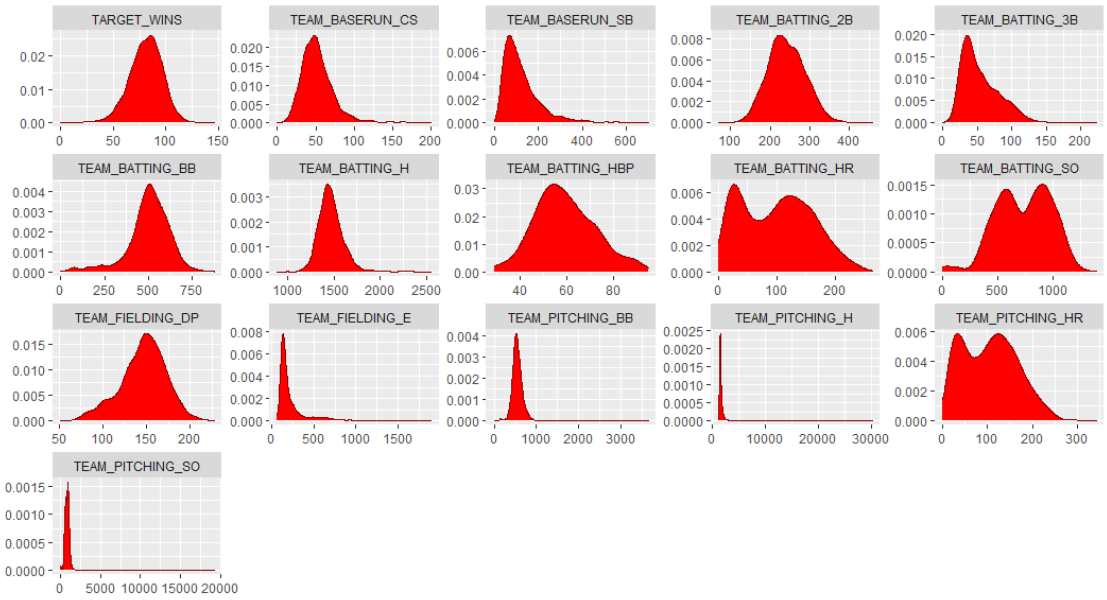
```
nrow(mb_train)

## [1] 2276

names(mb_train)

##  [1] "TARGET_WINS"      "TEAM_BATTING_H"   "TEAM_BATTING_2B"  "TEAM_BATTIN
G_3B"
##  [5] "TEAM_BATTING_HR"  "TEAM_BATTING_BB"  "TEAM_BATTING_SO"  "TEAM_BASERU
N_SB"
##  [9] "TEAM_BASERUN_CS"  "TEAM_BATTING_HBP" "TEAM_PITCHING_H"  "TEAM_PITCHI
NG_HR"
## [13] "TEAM_PITCHING_BB" "TEAM_PITCHING_SO" "TEAM_FIELDING_E"  "TEAM_FIELDI
NG_DP"

summary(mb_train)

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

```
##  Median : 536.5   Median :  813.5   Median : 159.0   Median :149.0
##  Mean   : 553.0   Mean   :  817.7   Mean   : 246.5   Mean   :146.4
##  3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2   3rd Qu.:164.0
##  Max.   :3645.0   Max.   :19278.0   Max.   :1898.0   Max.   :228.0
##                   NA's   :102                        NA's   :286

head(mb_train, n=5)

##   TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_
HR
## 1          39           1445             194              39
13
## 2          70           1339             219              22             1
90
## 3          86           1377             232              35             1
37
## 4          70           1387             209              38
96
## 5          82           1297             186              27             1
02
##   TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB TEAM_BASERUN_CS
## 1             143             842              NA              NA
## 2             685            1075              37              28
## 3             602             917              46              27
## 4             451             922              43              30
## 5             472             920              49              39
##   TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
## 1               NA            9364               84              927
## 2               NA            1347              191              689
## 3               NA            1377              137              602
## 4               NA            1396               97              454
## 5               NA            1297              102              472
##   TEAM_PITCHING_SO TEAM_FIELDING_E TEAM_FIELDING_DP
## 1             5456            1011               NA
## 2             1082             193              155
## 3              917             175              153
## 4              928             164              156
## 5              920             138              168
```

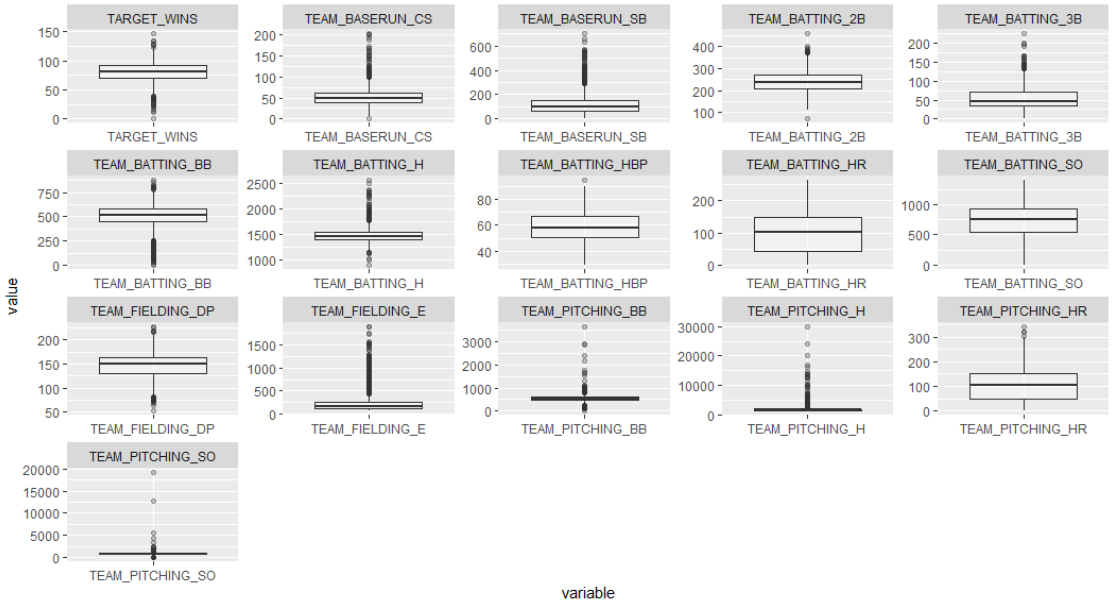
Missing Data

```
colSums(is.na(mb_train))

##      TARGET_WINS   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B
##                0                0                0                0
##  TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB
##                0                0              102              131
##  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H TEAM_PITCHING_HR
##              772             2085                0                0
## TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E TEAM_FIELDING_DP
##                0              102                0              286
```

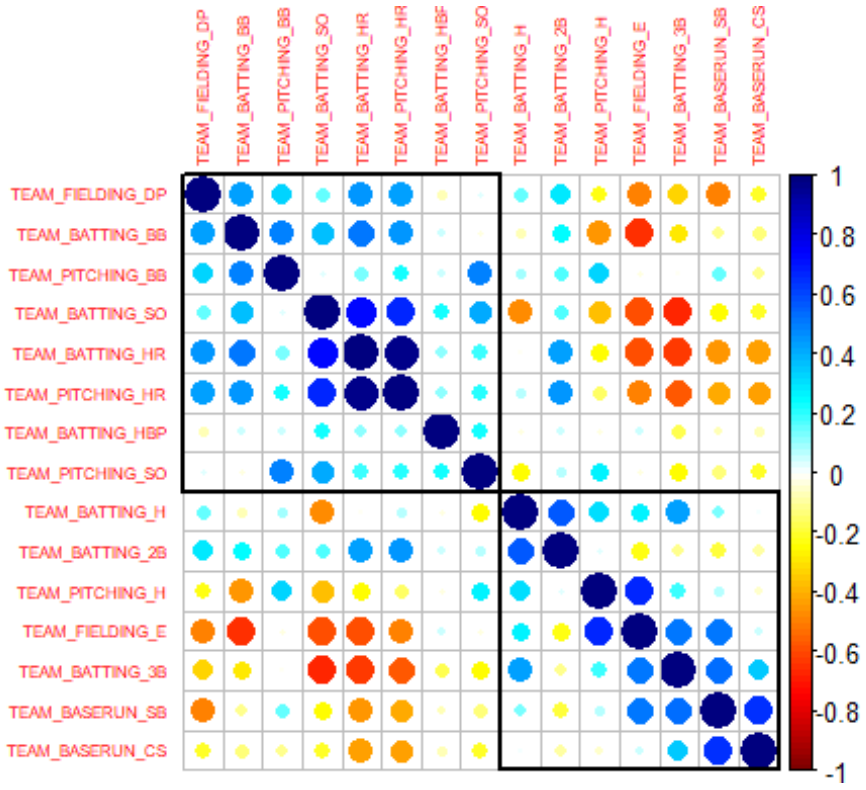
Density plot



Box plot



Corr Plot



Missing Data Percentage Calculation

```
col_sums_train <- colSums(mb_train %>% sapply(is.na))
pct_missing <- round(col_sums_train / nrow(mb_train) * 100, 2)
stack(sort(pct_missing, decreasing = TRUE))
```

```
##    values              ind
## 1   91.61 TEAM_BATTING_HBP
## 2   33.92  TEAM_BASERUN_CS
## 3   12.57 TEAM_FIELDING_DP
## 4    5.76  TEAM_BASERUN_SB
## 5    4.48  TEAM_BATTING_SO
## 6    4.48 TEAM_PITCHING_SO
## 7    0.00      TARGET_WINS
## 8    0.00   TEAM_BATTING_H
## 9    0.00  TEAM_BATTING_2B
## 10   0.00  TEAM_BATTING_3B
## 11   0.00  TEAM_BATTING_HR
## 12   0.00  TEAM_BATTING_BB
## 13   0.00  TEAM_PITCHING_H
## 14   0.00 TEAM_PITCHING_HR
## 15   0.00 TEAM_PITCHING_BB
## 16   0.00  TEAM_FIELDING_E
```

Data Exploration Next Steps

1. The below fields are having missing information in the descending order. TEAM_BATTING_HBP, TEAM_BASERUN_CS, TEAM_FIELDING_DP, TEAM_BASERUN_SB, TEAM_BATTING_SO, TEAM_PITCHING_SO
2. TEAM_BATTING_HBP need to be removed.
3. Set median values to TEAM_BASERUN_CS, TEAM_FIELDING_DP, TEAM_BASERUN_SB, TEAM_BATTING_SO, TEAM_PITCHING_SO

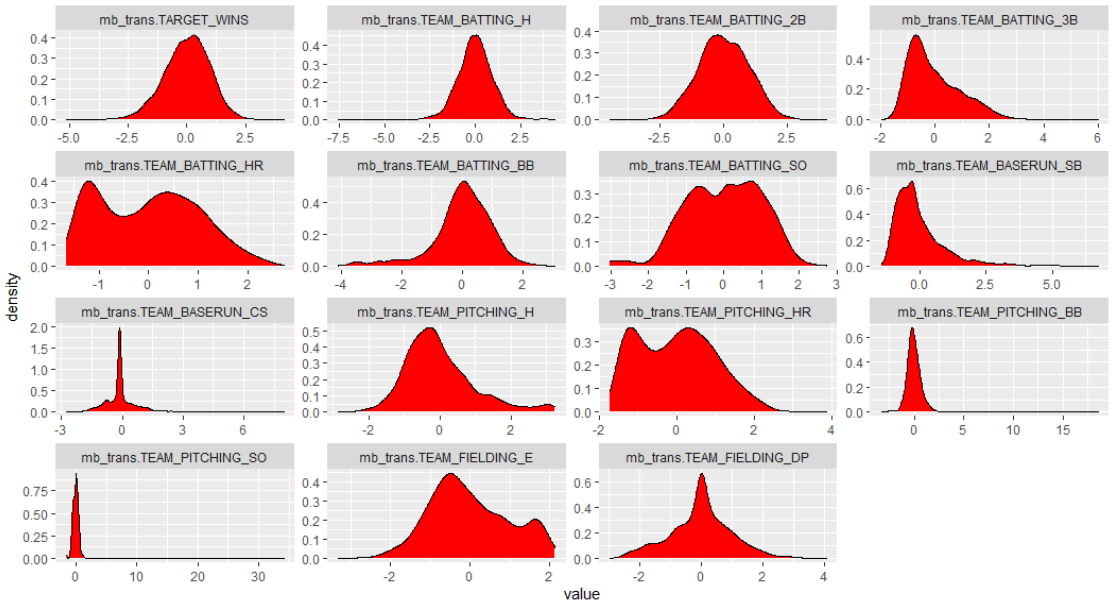
2. Data Preparation

Correct the missing information

```
mb_train <- mb_train %>%
   mutate_all(~ifelse(is.na(.), median(., na.rm = TRUE), .))
mb_train <- subset(mb_train, select = -c(TEAM_BATTING_HBP) )
```

Data Transformation

In the data transformation phase , Box Cox, centering and scaling methods are used for performing data transformation of predictors. Density PLots are created once again to showcase the difference between the original data and the transformed data.

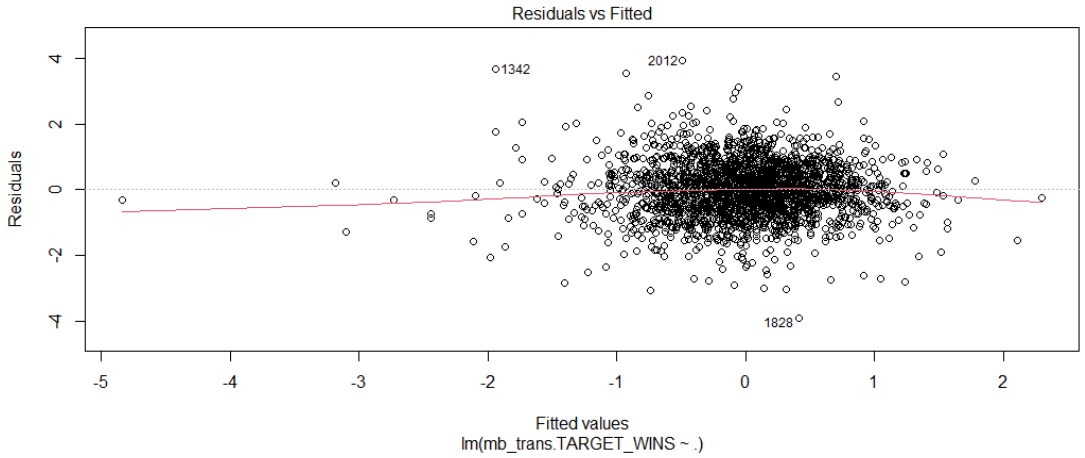


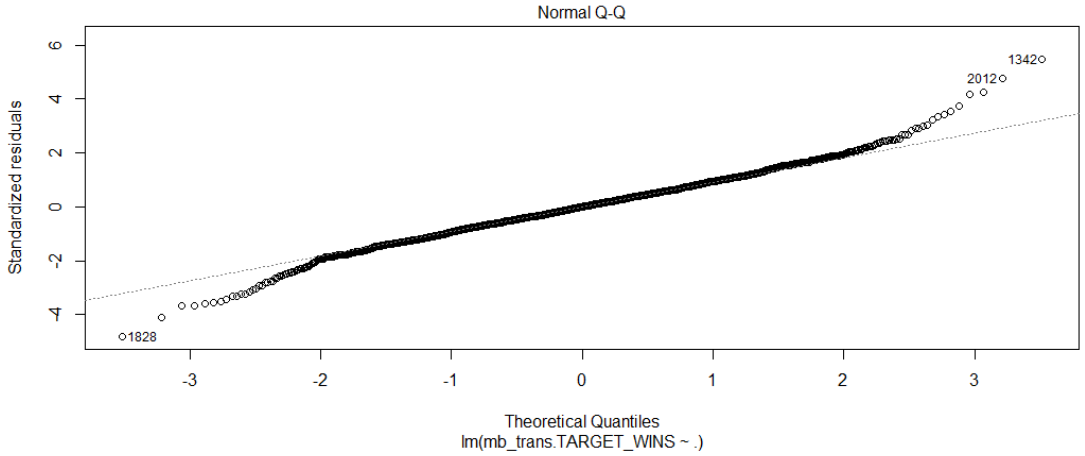
3. Models

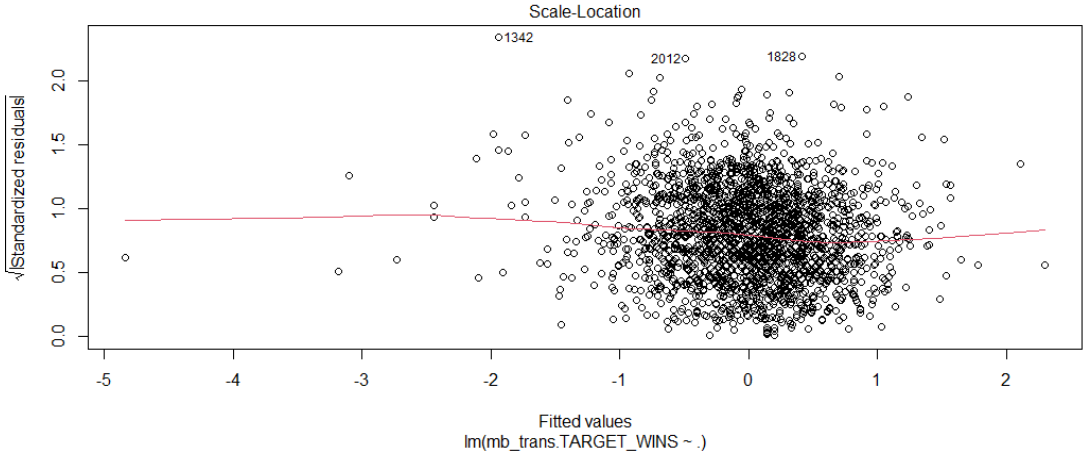
Linear Regression Model 1

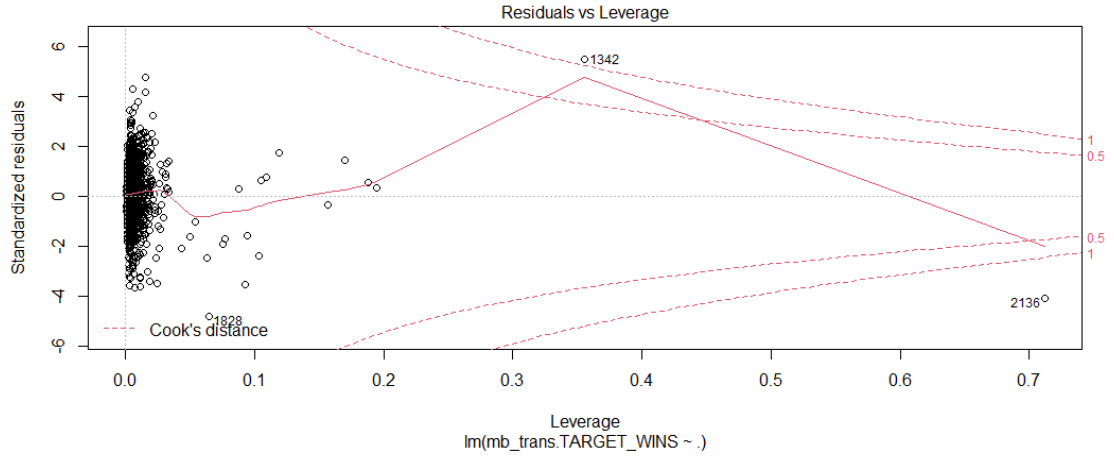
```
##
## Call:
## lm(formula = mb_trans.TARGET_WINS ~ ., data = mb_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.8923 -0.5159  0.0057  0.5110  3.9291
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.087e-12  1.748e-02   0.000  1.00000
## mb_trans.TEAM_BATTING_H     4.038e-01  3.774e-02  10.698  < 2e-16 ***
## mb_trans.TEAM_BATTING_2B   -7.106e-02  2.812e-02  -2.528  0.01156 *
## mb_trans.TEAM_BATTING_3B    2.053e-01  3.079e-02   6.668 3.26e-11 ***
## mb_trans.TEAM_BATTING_HR   -1.977e-02  1.115e-01  -0.177  0.85926
## mb_trans.TEAM_BATTING_BB    2.373e-01  3.813e-02   6.223 5.79e-10 ***
## mb_trans.TEAM_BATTING_SO   -1.943e-01  4.081e-02  -4.761 2.05e-06 ***
## mb_trans.TEAM_BASERUN_SB    1.461e-01  2.323e-02   6.290 3.79e-10 ***
## mb_trans.TEAM_BASERUN_CS   -8.221e-03  1.877e-02  -0.438  0.66134
## mb_trans.TEAM_PITCHING_H   -2.932e-02  4.082e-02  -0.718  0.47270
## mb_trans.TEAM_PITCHING_HR   1.718e-01  1.009e-01   1.702  0.08880 .
## mb_trans.TEAM_PITCHING_BB -1.197e-01  3.824e-02  -3.130  0.00177 **
## mb_trans.TEAM_PITCHING_SO   1.594e-01  3.221e-02   4.950 7.98e-07 ***
## mb_trans.TEAM_FIELDING_E   -3.672e-01  3.865e-02  -9.500  < 2e-16 ***
## mb_trans.TEAM_FIELDING_DP -1.973e-01  2.014e-02  -9.794  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8337 on 2261 degrees of freedom
## Multiple R-squared:  0.3092, Adjusted R-squared:  0.3049
## F-statistic: 72.27 on 14 and 2261 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
Fitted values
lm(mb_trans.TARGET_WINS ~ .)
1342
2012
1828


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(mb_trans.TARGET_WINS ~ .)
1342
2012
1828

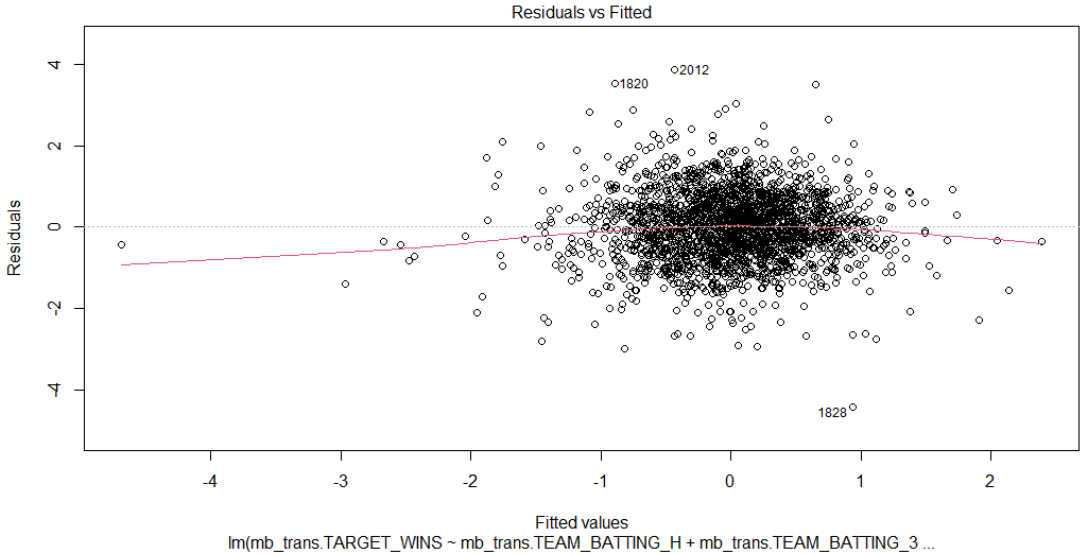


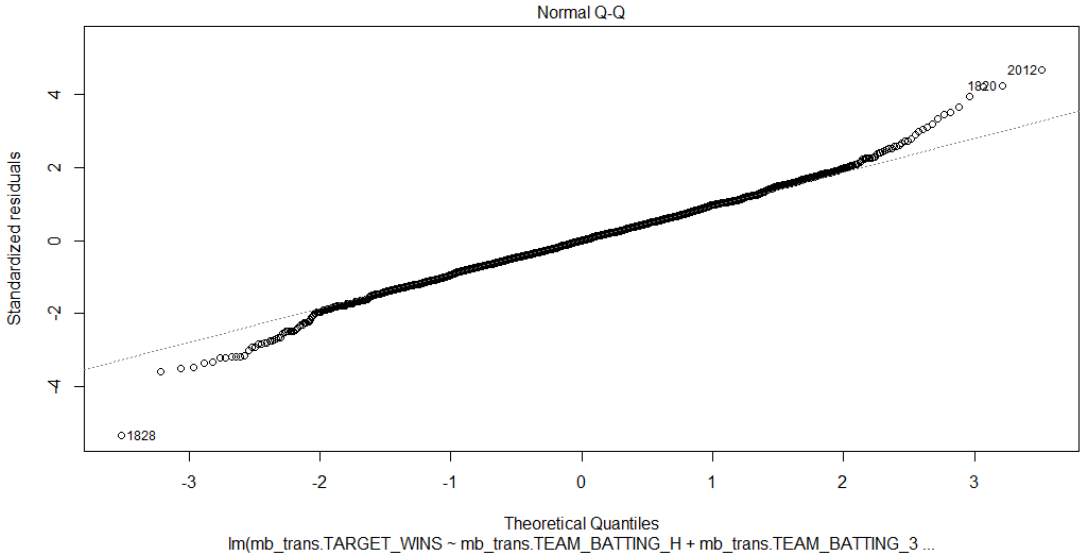


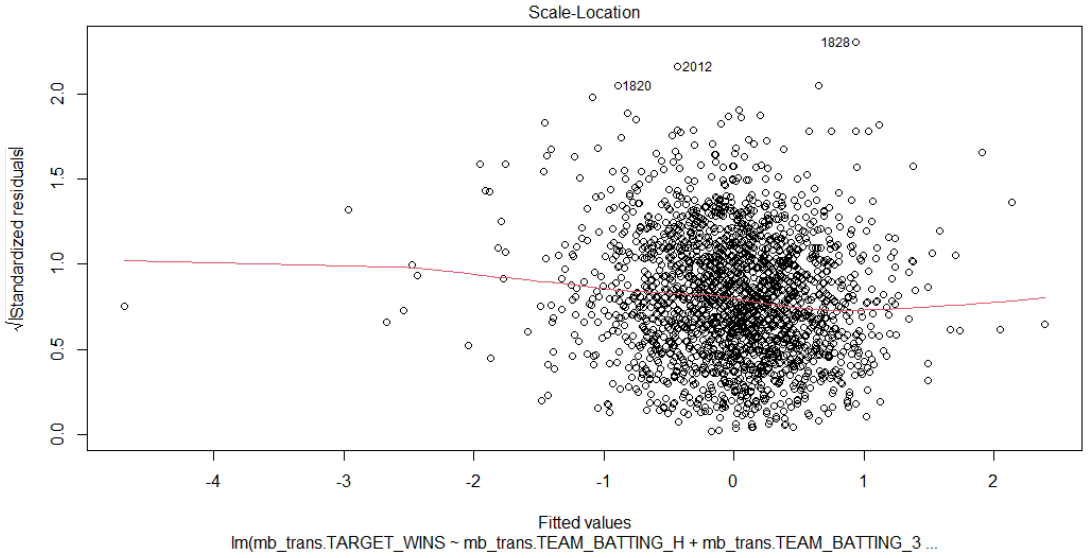
Linear Regression Model 2

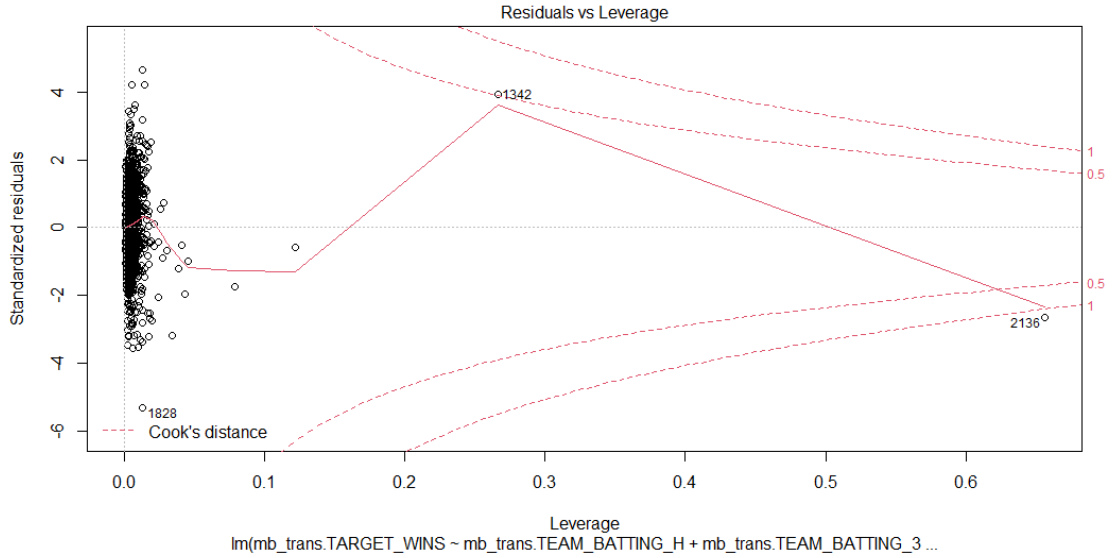
```
##
## Call:
## lm(formula = mb_trans.TARGET_WINS ~ mb_trans.TEAM_BATTING_H +
##     mb_trans.TEAM_BATTING_3B + mb_trans.TEAM_BATTING_HR + mb_trans.TEAM_BA
TTING_BB +
##     mb_trans.TEAM_BATTING_SO + mb_trans.TEAM_BASERUN_SB + mb_trans.TEAM_PI
TCHING_SO +
##     mb_trans.TEAM_PITCHING_H + mb_trans.TEAM_PITCHING_SO + mb_trans.TEAM_F
IELDING_E +
##     mb_trans.TEAM_FIELDING_DP, data = mb_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4150 -0.5244  0.0010  0.5193  3.8725
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  5.322e-12  1.752e-02   0.000 1.000000
```

```
## mb_trans.TEAM_BATTING_H    3.572e-01  3.122e-02  11.441  < 2e-16 ***
## mb_trans.TEAM_BATTING_3B   2.182e-01  3.048e-02   7.161 1.08e-12 ***
## mb_trans.TEAM_BATTING_HR   1.656e-01  3.847e-02   4.305 1.74e-05 ***
## mb_trans.TEAM_BATTING_BB   1.383e-01  2.414e-02   5.729 1.15e-08 ***
## mb_trans.TEAM_BATTING_SO  -1.678e-01  3.792e-02  -4.425 1.01e-05 ***
## mb_trans.TEAM_BASERUN_SB   1.409e-01  2.246e-02   6.276 4.14e-10 ***
## mb_trans.TEAM_PITCHING_SO  7.884e-02  2.237e-02   3.525 0.000431 ***
## mb_trans.TEAM_PITCHING_H  -5.018e-02  3.518e-02  -1.426 0.153925
## mb_trans.TEAM_FIELDING_E  -3.467e-01  3.757e-02  -9.227  < 2e-16 ***
## mb_trans.TEAM_FIELDING_DP -1.982e-01  2.019e-02  -9.817  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.836 on 2265 degrees of freedom
## Multiple R-squared:  0.3042, Adjusted R-squared:  0.3011
## F-statistic:    99 on 10 and 2265 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
Fitted values
lm(mb_trans.TARGET_WINS ~ mb_trans.TEAM_BATTING_H + mb_trans.TEAM_BATTING_3 ...
2012
1820
1828


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(mb_trans.TARGET_WINS ~ mb_trans.TEAM_BATTING_H + mb_trans.TEAM_BATTING_3 ...
2012
1820
1828

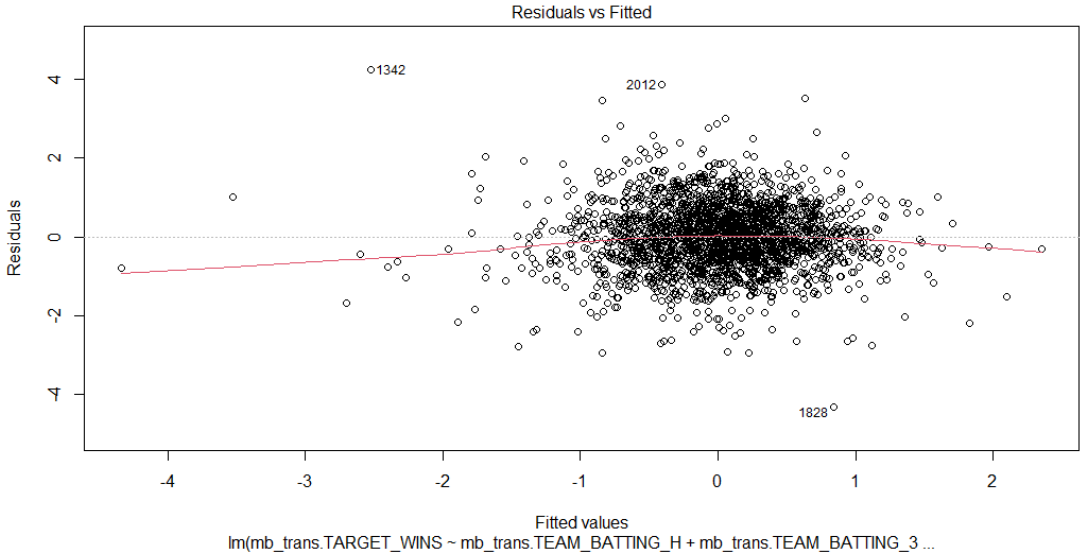


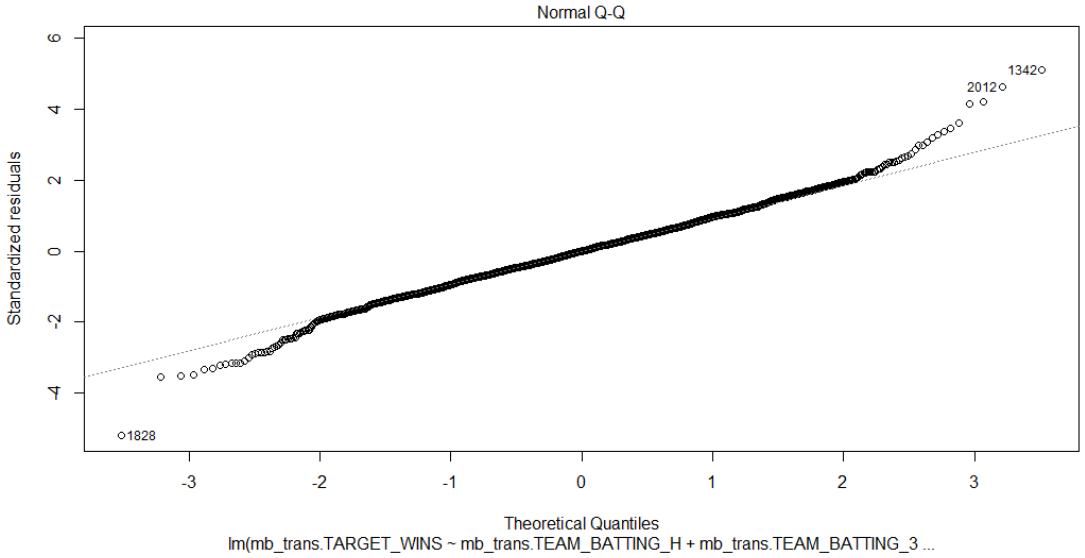


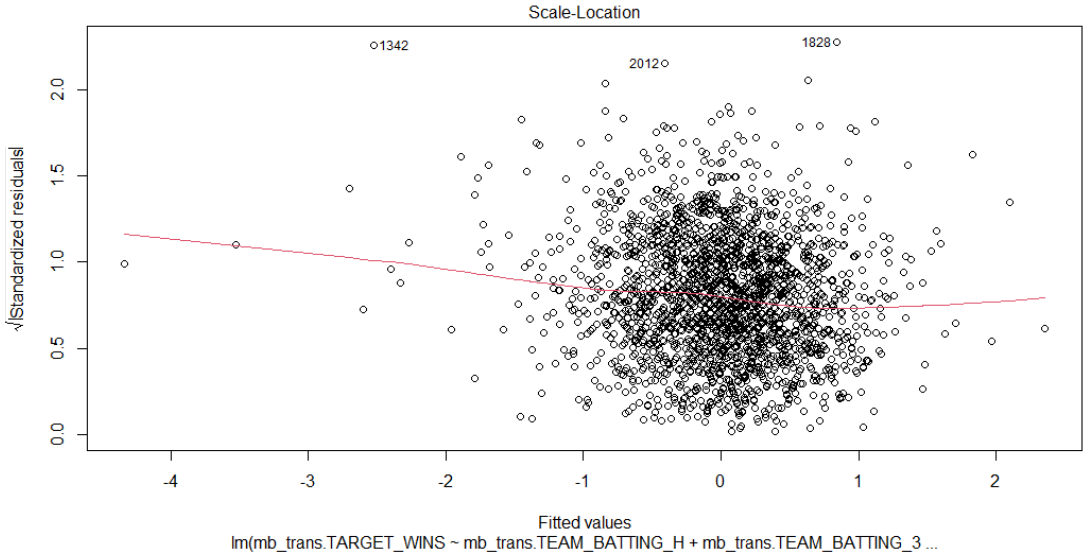
Linear Regression Model 3

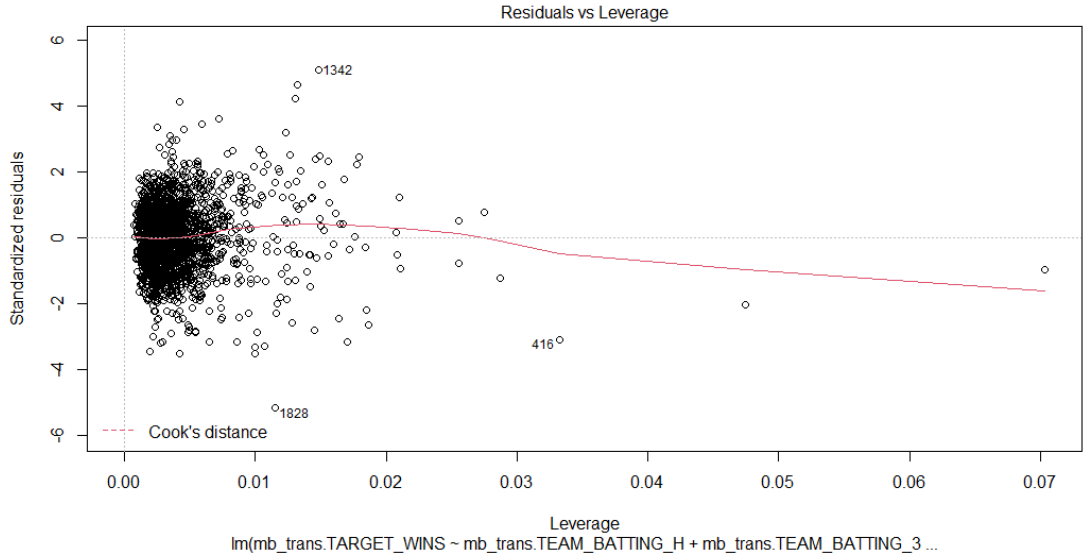
```
##
## Call:
## lm(formula = mb_trans.TARGET_WINS ~ mb_trans.TEAM_BATTING_H +
##     mb_trans.TEAM_BATTING_3B + mb_trans.TEAM_BATTING_HR + mb_trans.TEAM_BA
TTING_BB +
##     mb_trans.TEAM_BATTING_SO + mb_trans.TEAM_BASERUN_SB + mb_trans.TEAM_FI
ELDING_E +
##     mb_trans.TEAM_FIELDING_DP, data = mb_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3176 -0.5271  0.0019  0.5238  4.2521
```

```
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.643e-12  1.756e-02   0.000  1.00000
## mb_trans.TEAM_BATTING_H     3.282e-01  2.497e-02  13.145  < 2e-16 ***
## mb_trans.TEAM_BATTING_3B    2.149e-01  3.050e-02   7.043 2.48e-12 ***
## mb_trans.TEAM_BATTING_HR    1.545e-01  3.814e-02   4.052 5.24e-05 ***
## mb_trans.TEAM_BATTING_BB    1.420e-01  2.261e-02   6.280 4.04e-10 ***
## mb_trans.TEAM_BATTING_SO   -1.113e-01  3.400e-02  -3.272  0.00108 **
## mb_trans.TEAM_BASERUN_SB    1.370e-01  2.244e-02   6.105 1.20e-09 ***
## mb_trans.TEAM_FIELDING_E   -3.444e-01  3.558e-02  -9.678  < 2e-16 ***
## mb_trans.TEAM_FIELDING_DP  -1.950e-01  2.016e-02  -9.673  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8379 on 2267 degrees of freedom
## Multiple R-squared:  0.3003, Adjusted R-squared:  0.2979
## F-statistic: 121.6 on 8 and 2267 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
1342
2012
1828
Residuals
Fitted values
lm(mb_trans.TARGET_WINS ~ mb_trans.TEAM_BATTING_H + mb_trans.TEAM_BATTING_3 ...


Normal Q-Q
1342
2012
1828
Standardized residuals
Theoretical Quantiles
lm(mb_trans.TARGET_WINS ~ mb_trans.TEAM_BATTING_H + mb_trans.TEAM_BATTING_3 ...





4.Select Models

Model Metrics

```
Model <- c("Model 1", "Model 2", "Model 3")
Standard_Error <- c(0.8337, 0.836, 0.8379)
Multiple_R_squared <- c(0.3092, 0.3042, 0.3003)
Adjusted_R_squared <- c(0.3049, 0.3011, 0.2979)

df1 <- data.frame(Model, Standard_Error, Multiple_R_squared, Adjusted_R_squar
ed)
df1
```

```
##     Model Standard_Error Multiple_R_squared Adjusted_R_squared
## 1 Model 1         0.8337             0.3092             0.3049
## 2 Model 2         0.8360             0.3042             0.3011
## 3 Model 3         0.8379             0.3003             0.2979
```

ANOVA Test

```
anova(model1, model2, model3)

## Analysis of Variance Table
##
## Model 1: mb_trans.TARGET_WINS ~ mb_trans.TEAM_BATTING_H + mb_trans.TEAM_BA
TTING_2B +
##     mb_trans.TEAM_BATTING_3B + mb_trans.TEAM_BATTING_HR + mb_trans.TEAM_BA
TTING_BB +
##     mb_trans.TEAM_BATTING_SO + mb_trans.TEAM_BASERUN_SB + mb_trans.TEAM_BA
SERUN_CS +
##     mb_trans.TEAM_PITCHING_H + mb_trans.TEAM_PITCHING_HR + mb_trans.TEAM_P
ITCHING_BB +
##     mb_trans.TEAM_PITCHING_SO + mb_trans.TEAM_FIELDING_E + mb_trans.TEAM_F
IELDING_DP
## Model 2: mb_trans.TARGET_WINS ~ mb_trans.TEAM_BATTING_H + mb_trans.TEAM_BA
TTING_3B +
##     mb_trans.TEAM_BATTING_HR + mb_trans.TEAM_BATTING_BB + mb_trans.TEAM_BA
TTING_SO +
##     mb_trans.TEAM_BASERUN_SB + mb_trans.TEAM_PITCHING_SO + mb_trans.TEAM_P
ITCHING_H +
##     mb_trans.TEAM_PITCHING_SO + mb_trans.TEAM_FIELDING_E + mb_trans.TEAM_F
IELDING_DP
## Model 3: mb_trans.TARGET_WINS ~ mb_trans.TEAM_BATTING_H + mb_trans.TEAM_BA
TTING_3B +
##     mb_trans.TEAM_BATTING_HR + mb_trans.TEAM_BATTING_BB + mb_trans.TEAM_BA
TTING_SO +
##     mb_trans.TEAM_BASERUN_SB + mb_trans.TEAM_FIELDING_E + mb_trans.TEAM_FI
ELDING_DP
##   Res.Df    RSS Df Sum of Sq     F   Pr(>F)
## 1   2261 1571.7
## 2   2265 1583.0 -4  -11.3666 4.088 0.002642 **
## 3   2267 1591.8 -2   -8.6975 6.256 0.001952 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

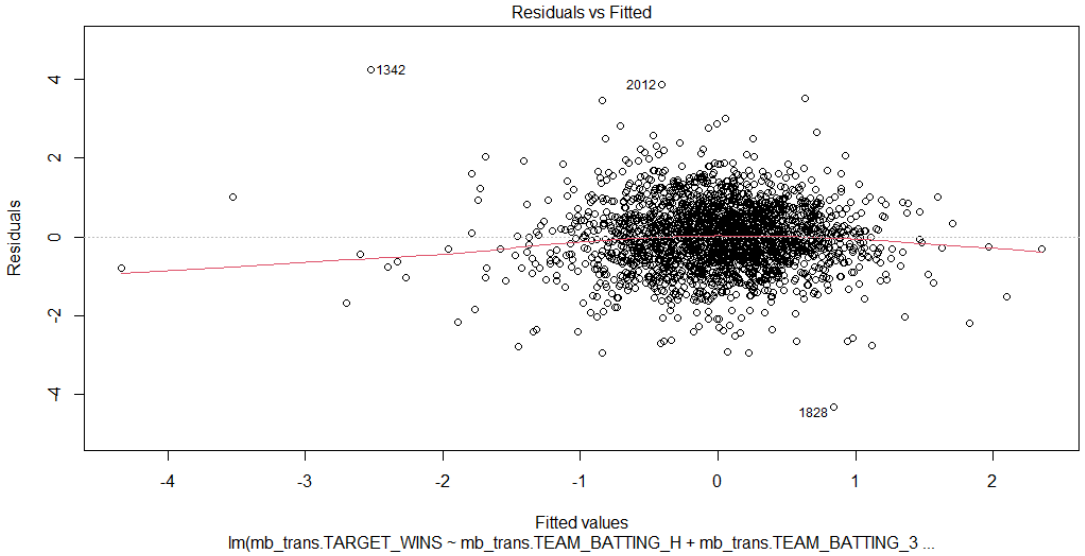
Conclusion

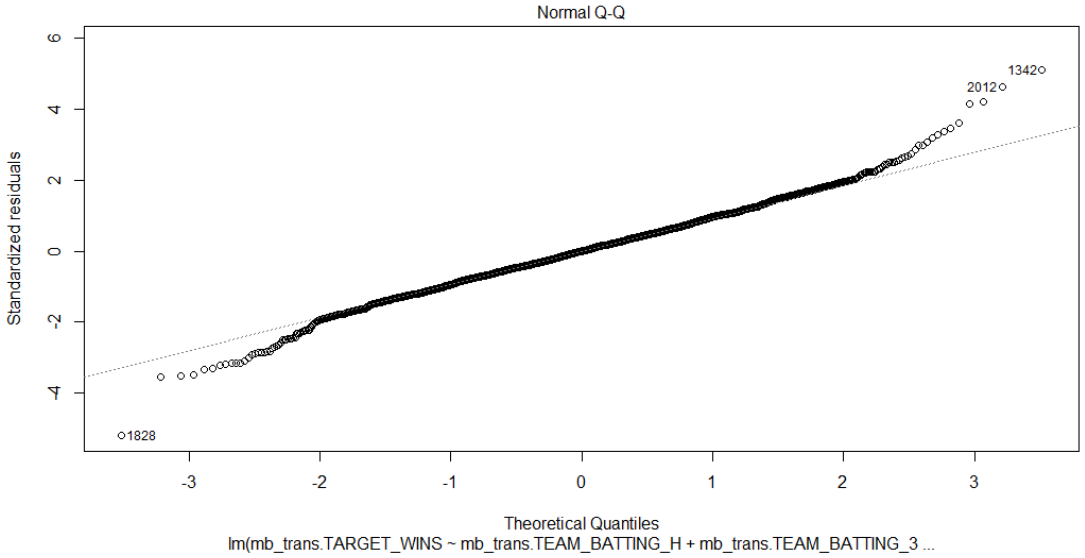
To Conclude , after comparing three models based on full dataset , partial dataset and key variables in the dataset , it is evident that the Linear regression model # 3 functions well as it satisfies the assumptions of linear regression and the p-value is very low.

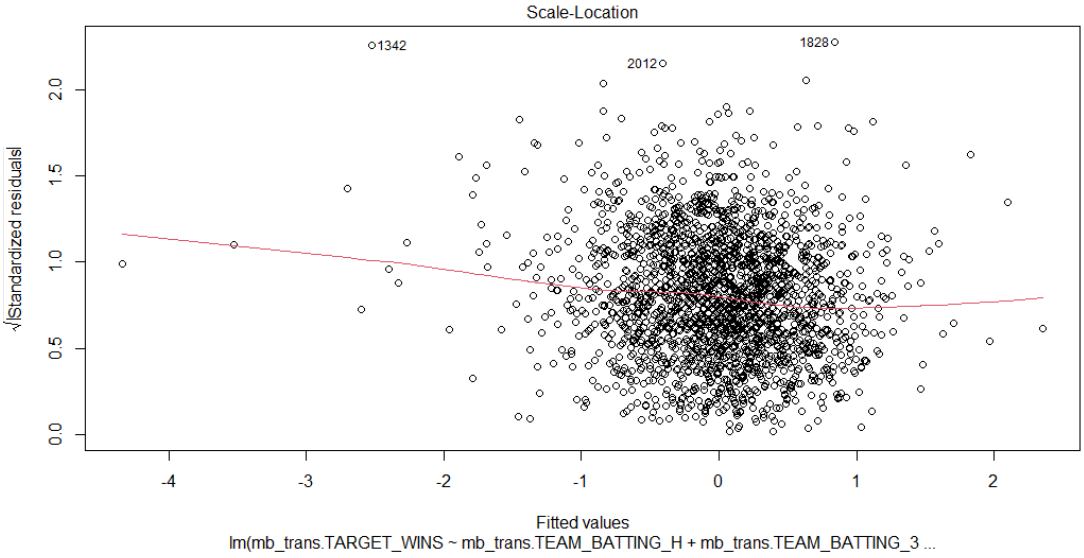
Hence Model selected is Linear Regression Model # 3.

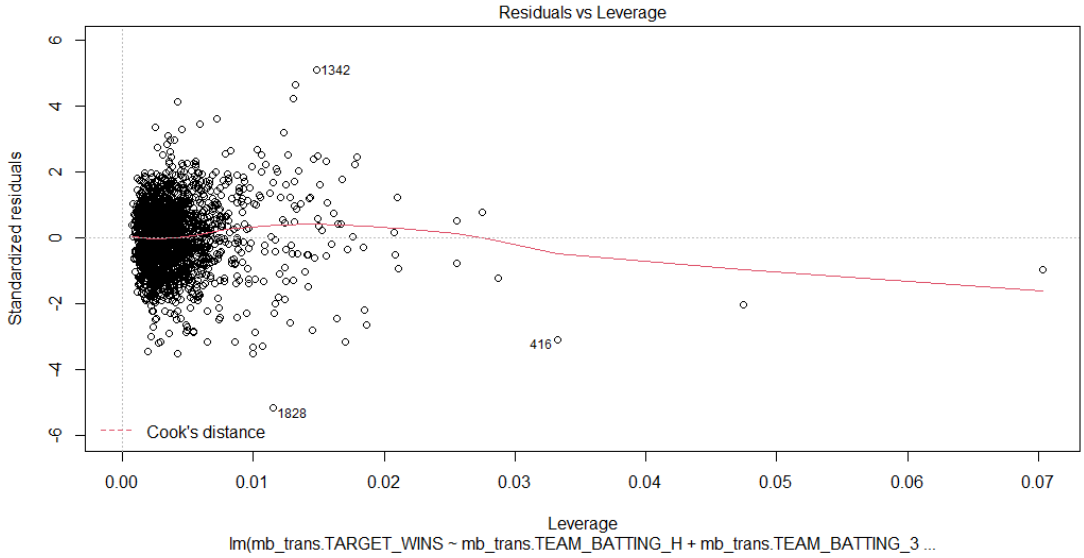
```
##
## Call:
```

```
## lm(formula = mb_trans.TARGET_WINS ~ mb_trans.TEAM_BATTING_H +
##     mb_trans.TEAM_BATTING_3B + mb_trans.TEAM_BATTING_HR + mb_trans.TEAM_BA
TTING_BB +
##     mb_trans.TEAM_BATTING_SO + mb_trans.TEAM_BASERUN_SB + mb_trans.TEAM_FI
ELDING_E +
##     mb_trans.TEAM_FIELDING_DP, data = mb_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3176 -0.5271  0.0019  0.5238  4.2521
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.643e-12  1.756e-02   0.000  1.00000
## mb_trans.TEAM_BATTING_H     3.282e-01  2.497e-02  13.145  < 2e-16 ***
## mb_trans.TEAM_BATTING_3B    2.149e-01  3.050e-02   7.043 2.48e-12 ***
## mb_trans.TEAM_BATTING_HR    1.545e-01  3.814e-02   4.052 5.24e-05 ***
## mb_trans.TEAM_BATTING_BB    1.420e-01  2.261e-02   6.280 4.04e-10 ***
## mb_trans.TEAM_BATTING_SO   -1.113e-01  3.400e-02  -3.272  0.00108 **
## mb_trans.TEAM_BASERUN_SB    1.370e-01  2.244e-02   6.105 1.20e-09 ***
## mb_trans.TEAM_FIELDING_E   -3.444e-01  3.558e-02  -9.678  < 2e-16 ***
## mb_trans.TEAM_FIELDING_DP -1.950e-01  2.016e-02  -9.673  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8379 on 2267 degrees of freedom
## Multiple R-squared:  0.3003, Adjusted R-squared:  0.2979
## F-statistic: 121.6 on 8 and 2267 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
1342
2012
1828
Residuals
Fitted values
lm(mb_trans.TARGET_WINS ~ mb_trans.TEAM_BATTING_H + mb_trans.TEAM_BATTING_3 ...


Normal Q-Q
1342
2012
1828
Standardized residuals
Theoretical Quantiles
lm(mb_trans.TARGET_WINS ~ mb_trans.TEAM_BATTING_H + mb_trans.TEAM_BATTING_3 ...





Appendix

https://github.com/jey1987/DATA621/blob/master/HW1/moneyball-model.rmd