SenthilDhanapal_IS621_HW1

Senthil Dhanapal

September 24, 2016

Variables Considered

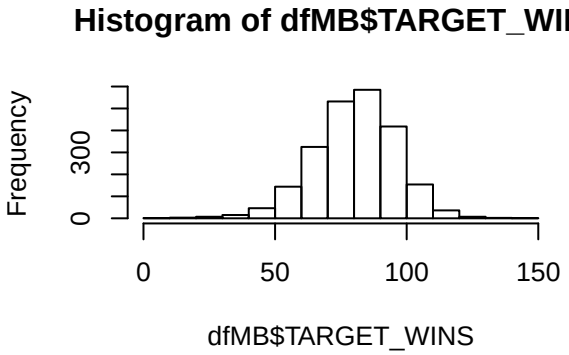
TEAM_BATTING_H = Base Hits by batters (1B,2B,3B,HR)

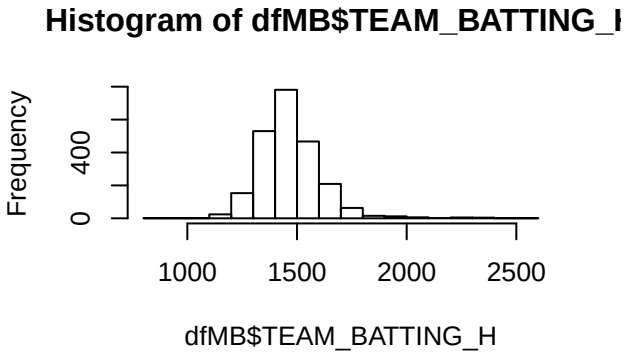
TEAM_BATTING_3B = Triples by batters (3B)

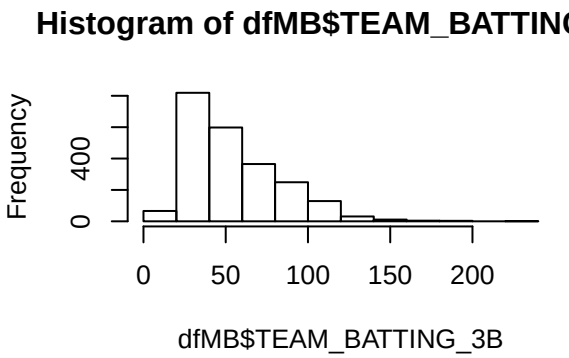
TEAM_PITCHING_H = Hits allowed

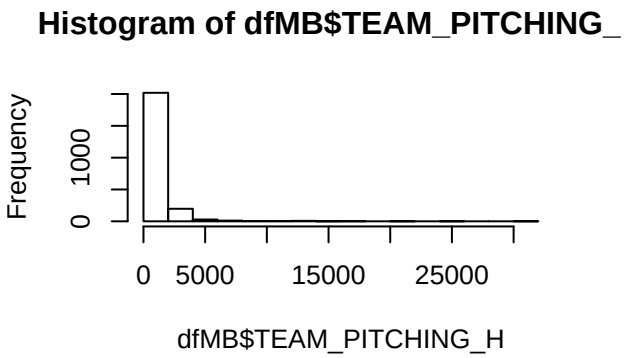
Data Exploration using summary and plots

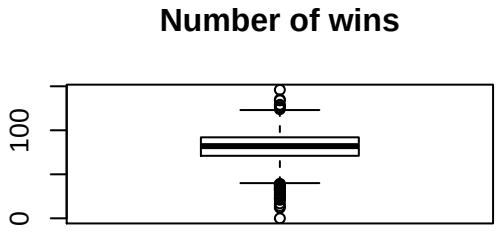
```
##   TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_3B  TEAM_PITCHING_H
##  Min.   :  0.00   Min.   : 891   Min.   :  0.00   Min.   : 1137
##  1st Qu.: 71.00   1st Qu.:1383   1st Qu.: 34.00   1st Qu.: 1419
##  Median : 82.00   Median :1454   Median : 47.00   Median : 1518
##  Mean   : 80.79   Mean   :1469   Mean   : 55.25   Mean   : 1779
##  3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.: 72.00   3rd Qu.: 1682
##  Max.   :146.00   Max.   :2554   Max.   :223.00   Max.   :30132
```

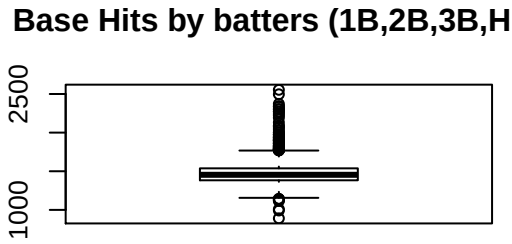

Histogram of dfMB$TARGET_WINS
Frequency
0
300
0
50
100
150
dfMB$TARGET_WINS

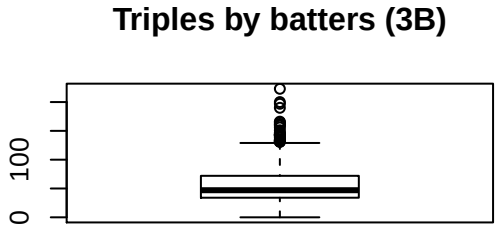

Histogram of dfMB$TEAM_BATTING_H
Frequency
0
400
1000
1500
2000
2500
dfMB$TEAM_BATTING_H

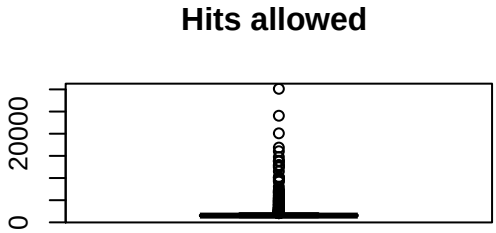

Histogram of dfMB$TEAM_BATTING_3
Frequency
0
400
0
50
100
150
200
dfMB$TEAM_BATTING_3B


Histogram of dfMB$TEAM_PITCHING_
Frequency
0
1000
0
5000
15000
25000
dfMB$TEAM_PITCHING_H


Number of wins
0
100

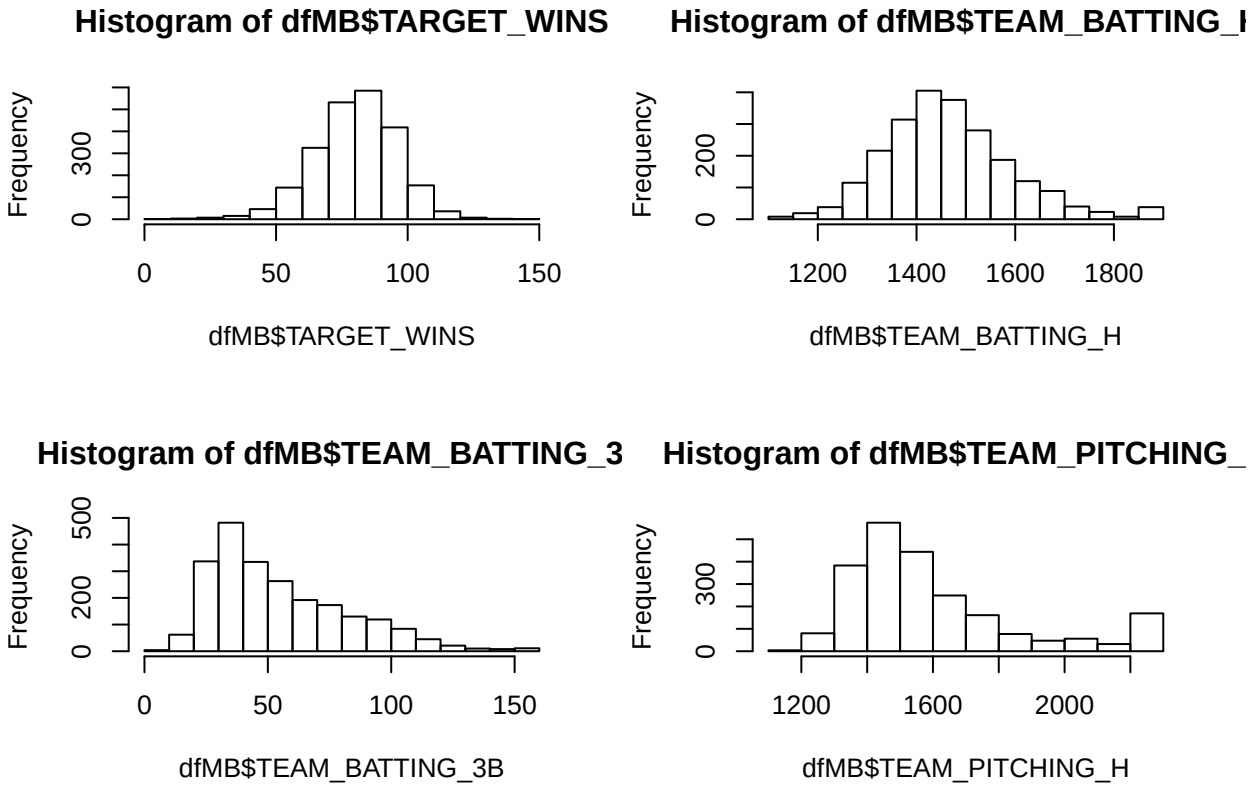

Base Hits by batters (1B,2B,3B,HR)
1000
2500

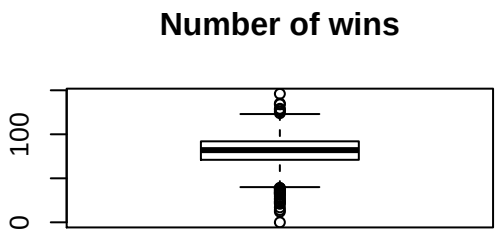

Triples by batters (3B)
0
100

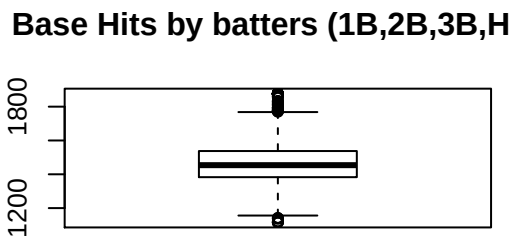

Hits allowed
0
20000

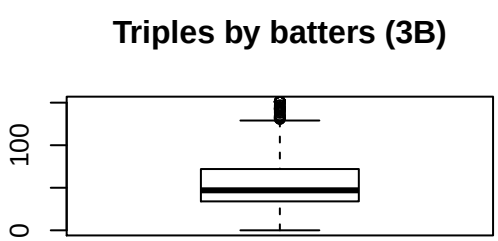
Handling Outliers using Winsorizing method

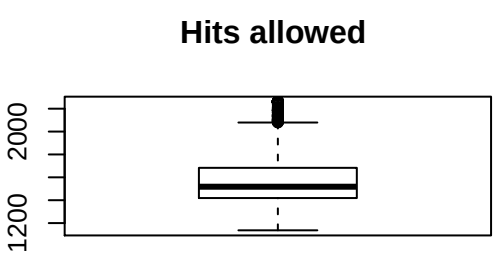
Data Exploration after winsorizing using plots







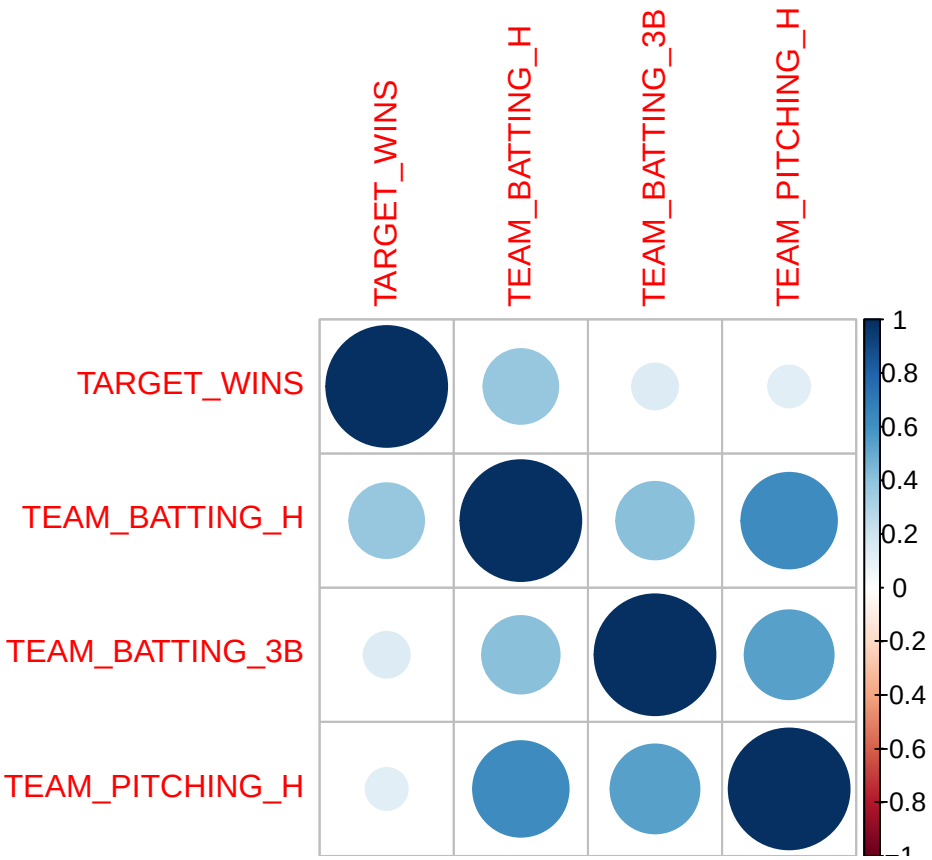




Check for relevancy and multicollinearity using Correllation between all variables

---

```
##                 TARGET_WINS TEAM_BATTING_H TEAM_BATTING_3B TEAM_PITCHING_H
## TARGET_WINS       1.0000000      0.3827511       0.1443922       0.1208252
## TEAM_BATTING_H    0.3827511      1.0000000       0.4114424       0.6256801
## TEAM_BATTING_3B   0.1443922      0.4114424       1.0000000       0.5412845
## TEAM_PITCHING_H   0.1208252      0.6256801       0.5412845       1.0000000
```



Correllation betwen TEAM_BATTING_H and TEAM_BATTING_3B may lead to multicollinearity

TEAM_BATTING_H has the best correllation with Target wins than any other predictors.

TEAM_PITCHING_H has no correllation

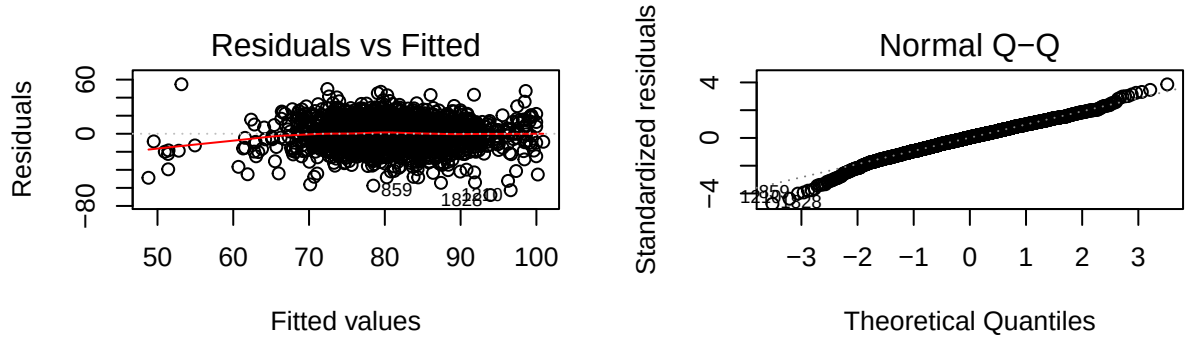
However, keeping all 3 variables to build models

Build Models

Model1

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_PITCHING_H +
```

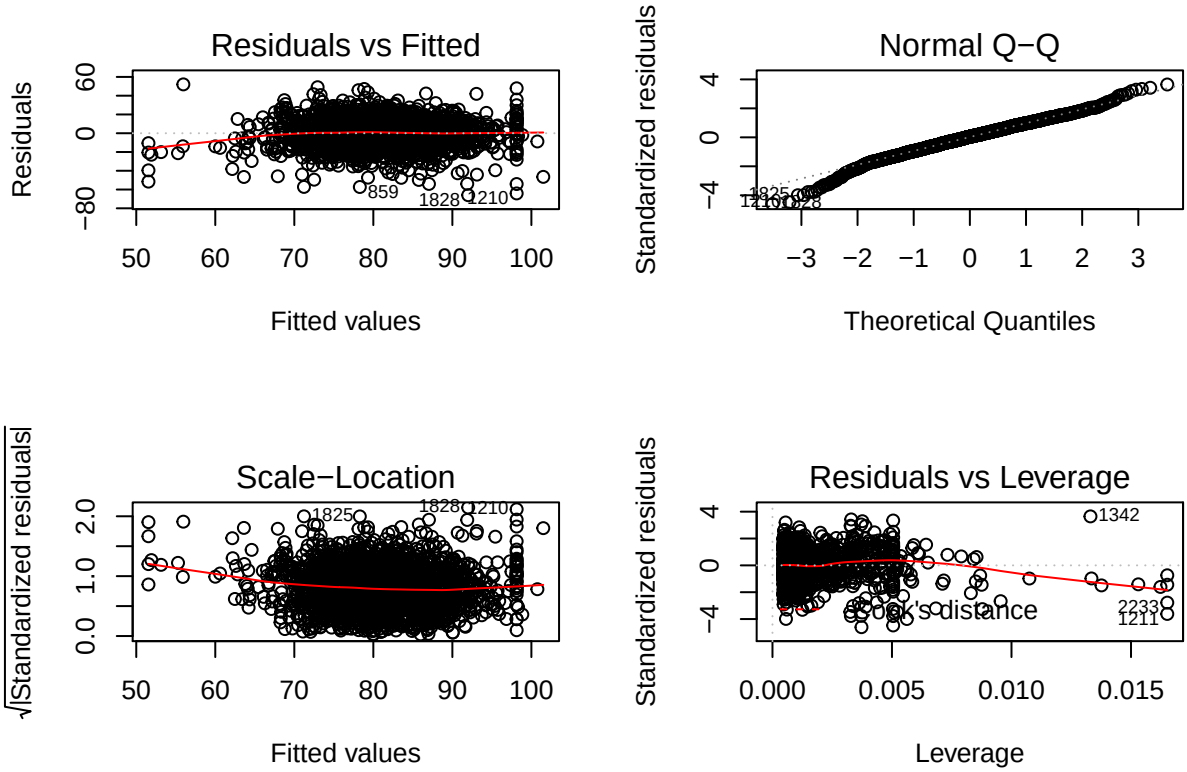
```
##     TEAM_BATTING_3B, data = dfMB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.937  -8.919   0.622   9.460  54.839
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     11.748390   3.521956   3.336 0.000864 ***
## TEAM_BATTING_H   0.060423   0.002990  20.207  < 2e-16 ***
## TEAM_PITCHING_H -0.013438   0.001607  -8.364  < 2e-16 ***
## TEAM_BATTING_3B  0.034748   0.013115   2.649 0.008120 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.34 on 2272 degrees of freedom
## Multiple R-squared:  0.1722, Adjusted R-squared:  0.1711
## F-statistic: 157.5 on 3 and 2272 DF,  p-value: < 2.2e-16
```



Model2

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_PITCHING_H,
```

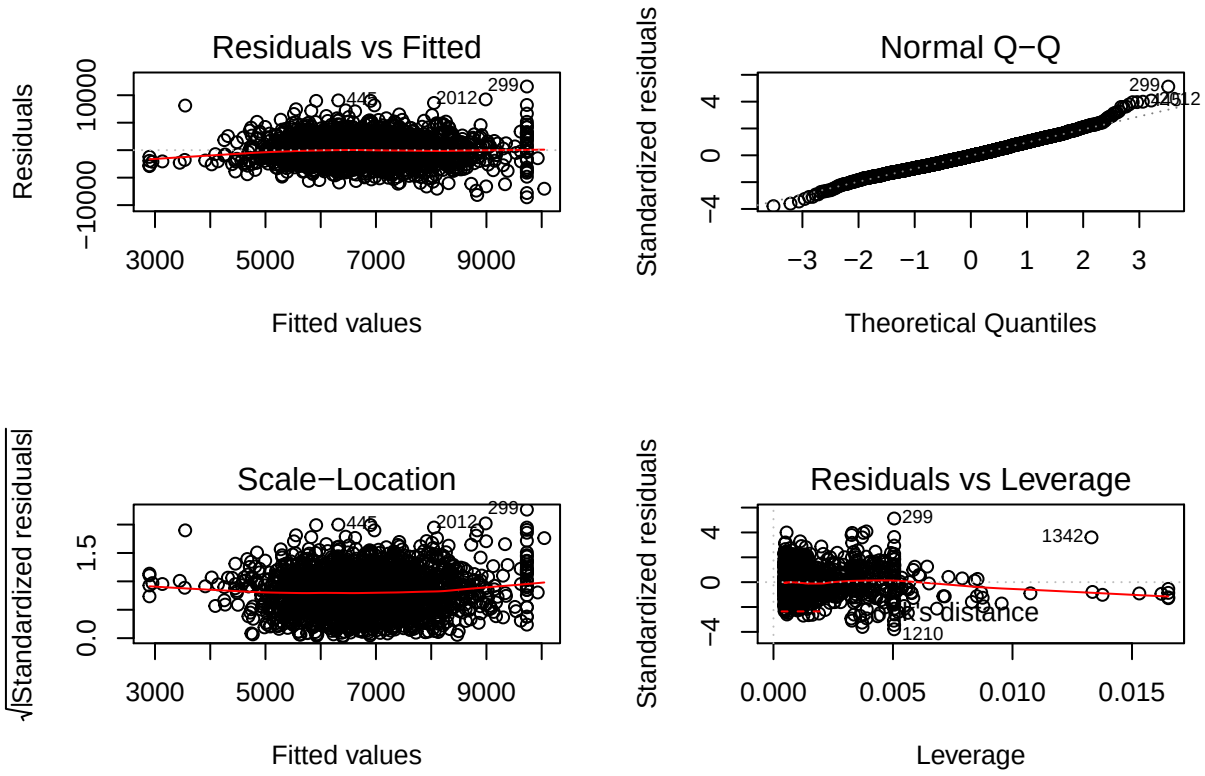
```
##     data = dfMB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.004  -8.969   0.613   9.542  52.042
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.660526   3.437200   2.811  0.00499 **
## TEAM_BATTING_H   0.061302   0.002976  20.601  < 2e-16 ***
## TEAM_PITCHING_H -0.011739   0.001475  -7.958 2.73e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.36 on 2273 degrees of freedom
## Multiple R-squared:  0.1696, Adjusted R-squared:  0.1689
## F-statistic: 232.2 on 2 and 2273 DF,  p-value: < 2.2e-16
```



Model3

```
##
## Call:
## lm(formula = TARGET_WINS^2 ~ TEAM_BATTING_H + TEAM_PITCHING_H,
##     data = dfMB)
```

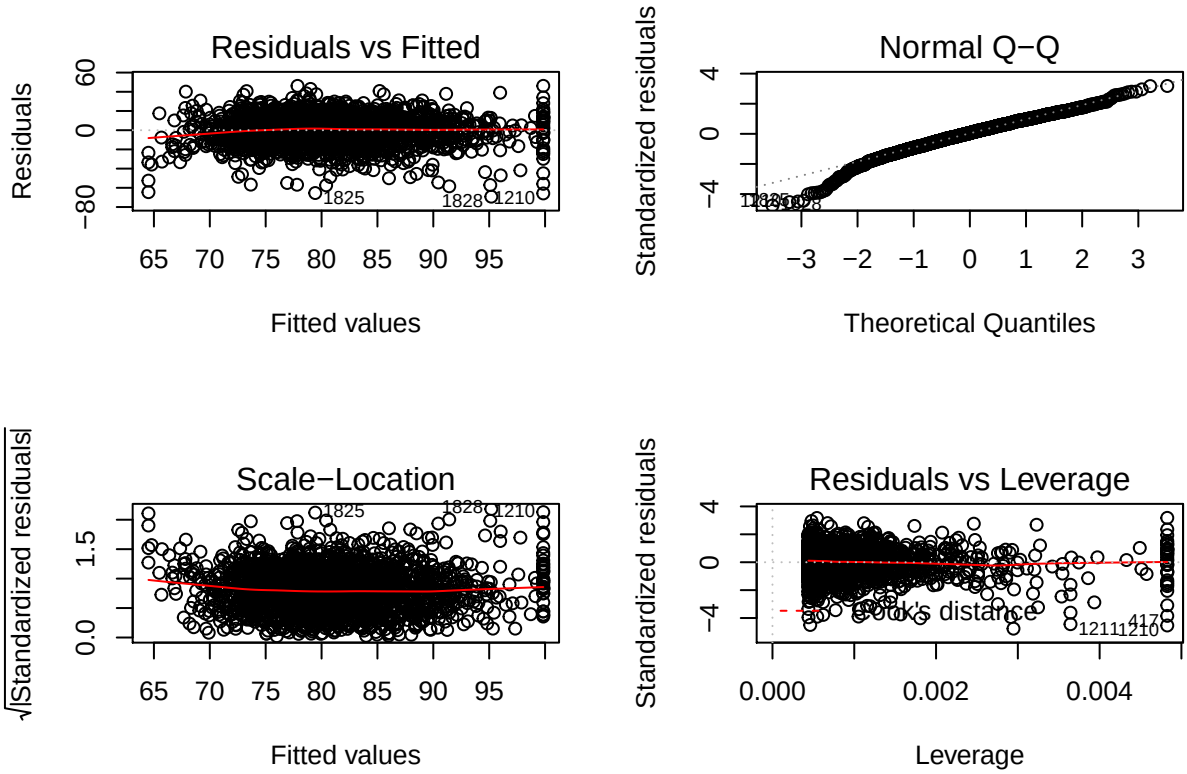
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8576.0 -1537.8  -101.1  1442.0 11584.0
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -4652.7368   542.7857  -8.572  < 2e-16 ***
## TEAM_BATTING_H      8.9899     0.4699  19.131  < 2e-16 ***
## TEAM_PITCHING_H    -1.0975     0.2329  -4.712 2.61e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2268 on 2273 degrees of freedom
## Multiple R-squared:  0.166,  Adjusted R-squared:  0.1653
## F-statistic: 226.3 on 2 and 2273 DF,  p-value: < 2.2e-16
```



Model4

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H, data = dfMB)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -69.191  -8.700   0.696   9.713  46.160
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12.633214   3.463354   3.648 0.000271 ***
## TEAM_BATTING_H  0.046485   0.002353  19.756  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.56 on 2274 degrees of freedom
## Multiple R-squared:  0.1465, Adjusted R-squared:  0.1461
## F-statistic: 390.3 on 1 and 2274 DF,  p-value: < 2.2e-16
```



Choosing Models

```
##   model adjRSqrd FStat PValForFStat ResidualVsFittedConstantVariation
## 1     1   0.1711 157.5  Significant                          Not good,
## 2     2   0.1689 232.2  Significant                           Not good
## 3     3   0.1653 226.3  Significant                           Not good
## 4     4   0.1461 390.3  Significant                           Not good
```

```
##   ResidualVsFittedCuve ResidualVsFittedCuveHeteroscedasticity
## 1         shows curve,                                    yes,
## 2          shows curve                                     yes
## 3          shows curve                                     yes
## 4          shows curve                                     yes
```