Critical Thinking Group 4 - HW1

Sreejaya, Suman, Vuthy

September 12, 2016

Purpose

The purpose of this experiment is to try to predict the amount of wins for a baseball team using the (modified) moneyball dataset. This dataset contains approximately 2200 observations with 17 variables. Each observation represents the performance of a professional baseball team from 1871 to 2006. The statistics have been adjusted to match the performance of a 162 game season.

Dataset:
Moneyball Training Data
Moneyball Evaluation Data

1. Data Exploration

The dependent (response) variable is *TARGET_WINS*. Excluding INDEX, the rest of the variables are the independent variables (predictors). Lets review how each of these independent variables are distributed & how each of these indepdent variable relates to the response variable 'TARGET_WINS'.

1.1 Missing Values

Review the *measure of the center* for the given variables. A quick look at the summary statistics indicate that there are missing values for some of the predictors.

```
##  TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B  TEAM_BATTING_HR
##  Min.   : 891   Min.   : 69.0   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00   1st Qu.: 42.00
##  Median :1454   Median :238.0   Median : 47.00   Median :102.00
##  Mean   :1469   Mean   :241.2   Mean   : 55.25   Mean   : 99.61
##  3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00   3rd Qu.:147.00
##  Max.   :2554   Max.   :458.0   Max.   :223.00   Max.   :264.00
##
##  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB TEAM_BASERUN_CS
##  Min.   :  0.0   Min.   :   0.0   Min.   :  0.0   Min.   :  0.0
##  1st Qu.:451.0   1st Qu.: 548.0   1st Qu.: 66.0   1st Qu.: 38.0
##  Median :512.0   Median : 750.0   Median :101.0   Median : 49.0
##  Mean   :501.6   Mean   : 735.6   Mean   :124.8   Mean   : 52.8
##  3rd Qu.:580.0   3rd Qu.: 930.0   3rd Qu.:156.0   3rd Qu.: 62.0
##  Max.   :878.0   Max.   :1399.0   Max.   :697.0   Max.   :201.0
##                  NA's   :102      NA's   :131     NA's   :772
##  TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
##  Min.   :29.00    Min.   : 1137   Min.   :  0.0    Min.   :   0.0
##  1st Qu.:50.50    1st Qu.: 1419   1st Qu.: 50.0    1st Qu.: 476.0
##  Median :58.00    Median : 1518   Median :107.0    Median : 536.5
##  Mean   :59.36    Mean   : 1779   Mean   :105.7    Mean   : 553.0
##  3rd Qu.:67.00    3rd Qu.: 1682   3rd Qu.:150.0    3rd Qu.: 611.0
```

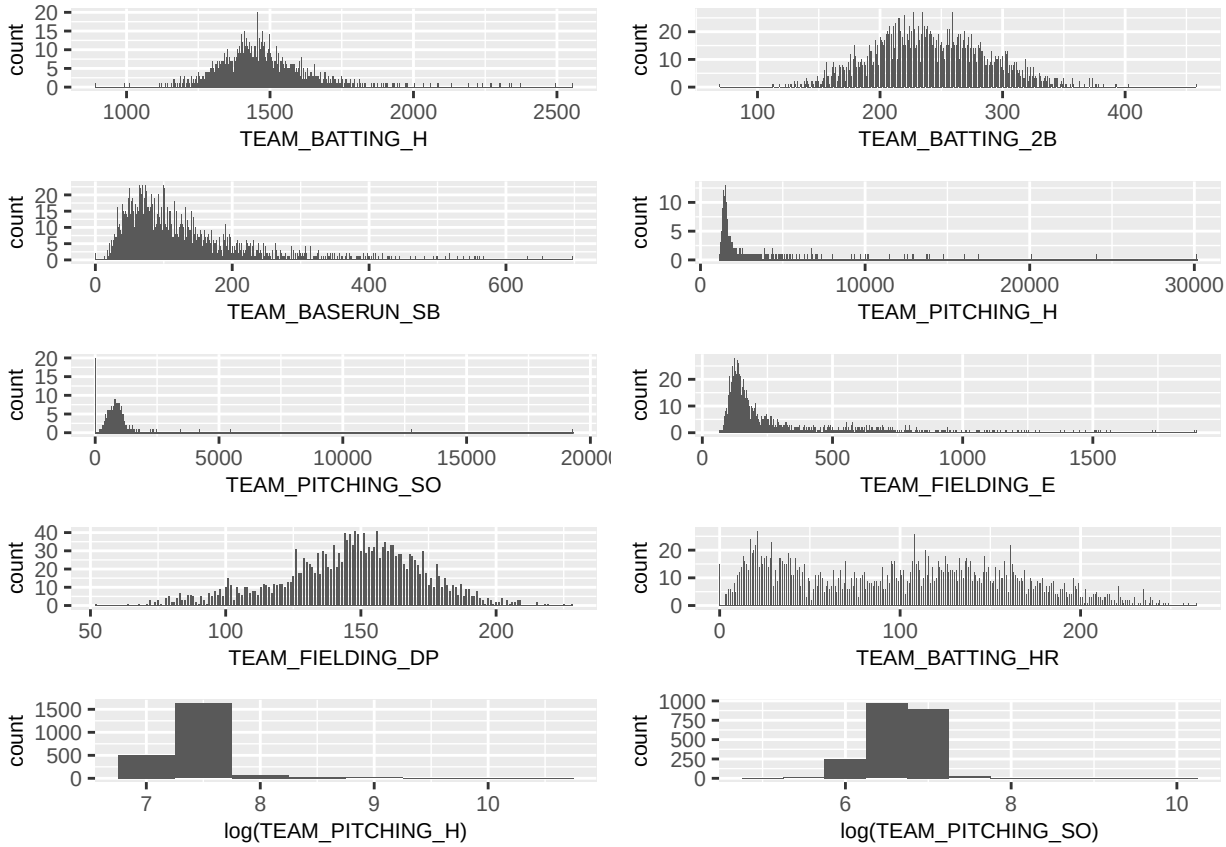
```
##  Max.   :95.00    Max.   :30132   Max.   :343.0    Max.   :3645.0
##  NA's   :2085
##  TEAM_PITCHING_SO  TEAM_FIELDING_E  TEAM_FIELDING_DP
##  Min.   :    0.0   Min.   :  65.0   Min.   : 52.0
##  1st Qu.:  615.0   1st Qu.: 127.0   1st Qu.:131.0
##  Median :  813.5   Median : 159.0   Median :149.0
##  Mean   :  817.7   Mean   : 246.5   Mean   :146.4
##  3rd Qu.:  968.0   3rd Qu.: 249.2   3rd Qu.:164.0
##  Max.   :19278.0   Max.   :1898.0   Max.   :228.0
##  NA's   :102                        NA's   :286
```

The list of predictor variables with missing data and their counts:

	Missing	Percentage
TEAM_BATTING_H	0	0.0000000
TEAM_BATTING_2B	0	0.0000000
TEAM_BATTING_3B	0	0.0000000
TEAM_BATTING_HR	0	0.0000000
TEAM_BATTING_BB	0	0.0000000
TEAM_BATTING_SO	102	0.0448155
TEAM_BASERUN_SB	131	0.0575571
TEAM_BASERUN_CS	772	0.3391916
TEAM_BATTING_HBP	2085	0.9160808
TEAM_PITCHING_H	0	0.0000000
TEAM_PITCHING_HR	0	0.0000000
TEAM_PITCHING_BB	0	0.0000000
TEAM_PITCHING_SO	102	0.0448155
TEAM_FIELDING_E	0	0.0000000
TEAM_FIELDING_DP	286	0.1256591

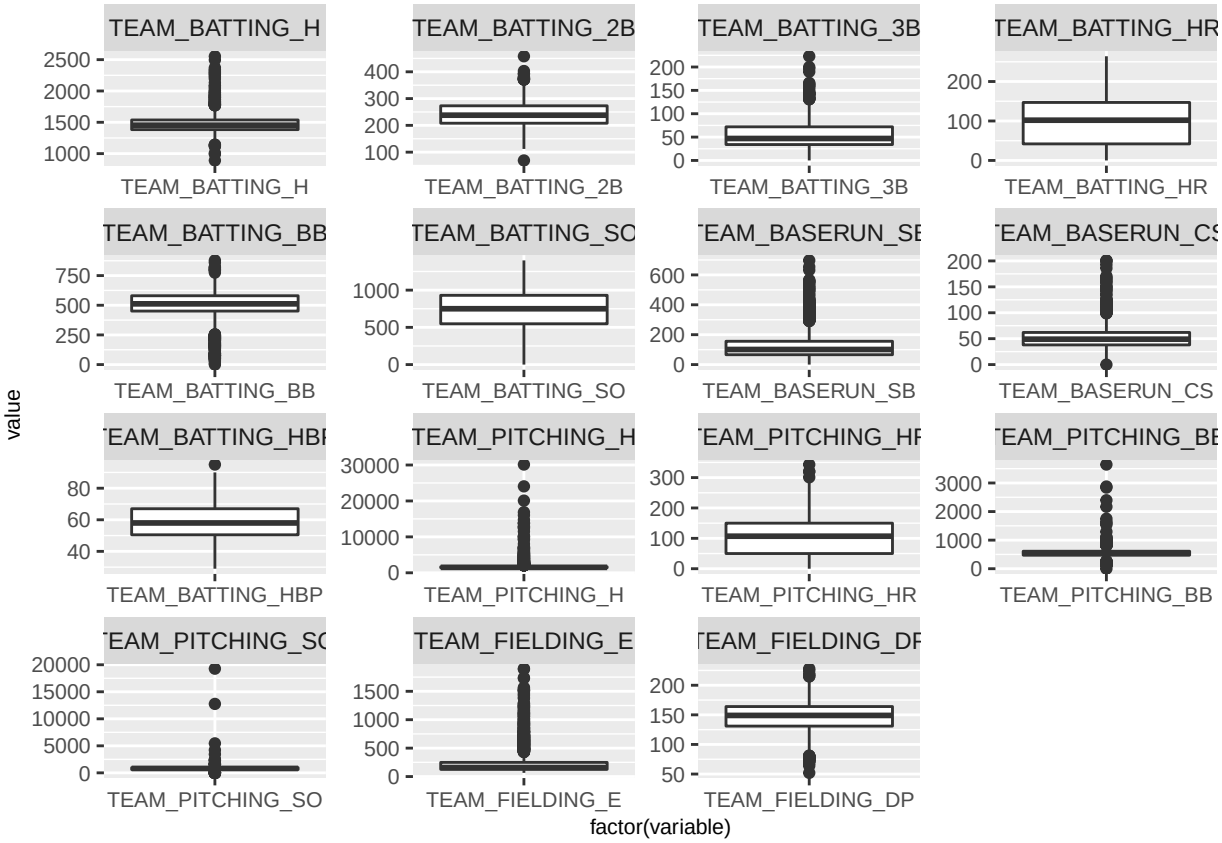
1.2 Distribution of predictors

Review the distributions of the predictors. Here are few histograms of the predictors.

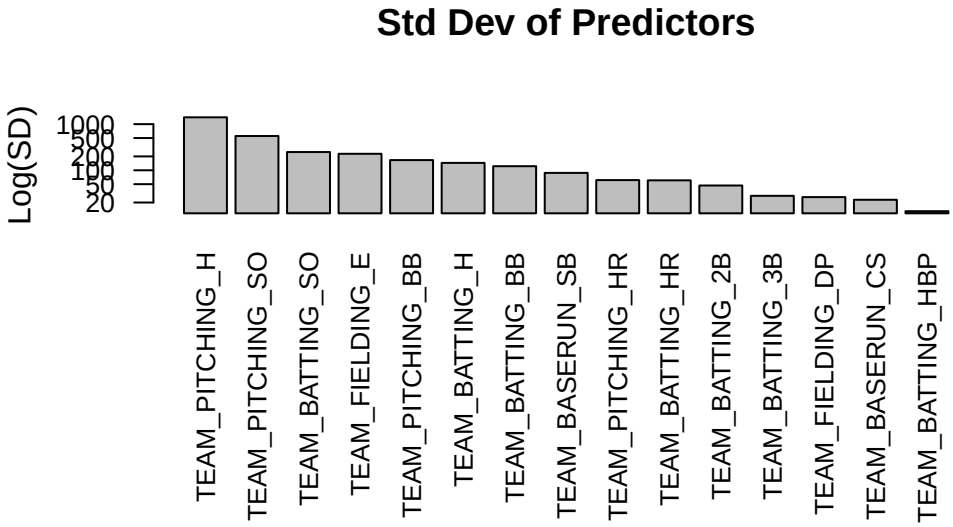


Based on the summary of the data, and the histograms, there are outliers and the distributions of the few of the predictors are skewed. Notice that *TEAM_PITCHING_H* and *TEAM_PICTCHING_SO* distributions are not visible at all in the above diagram, so the log transformation has been applied in the above.

Lets also review the box plot's of the predictors.


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
value
factor(variable)

1.3 Standard Deviation



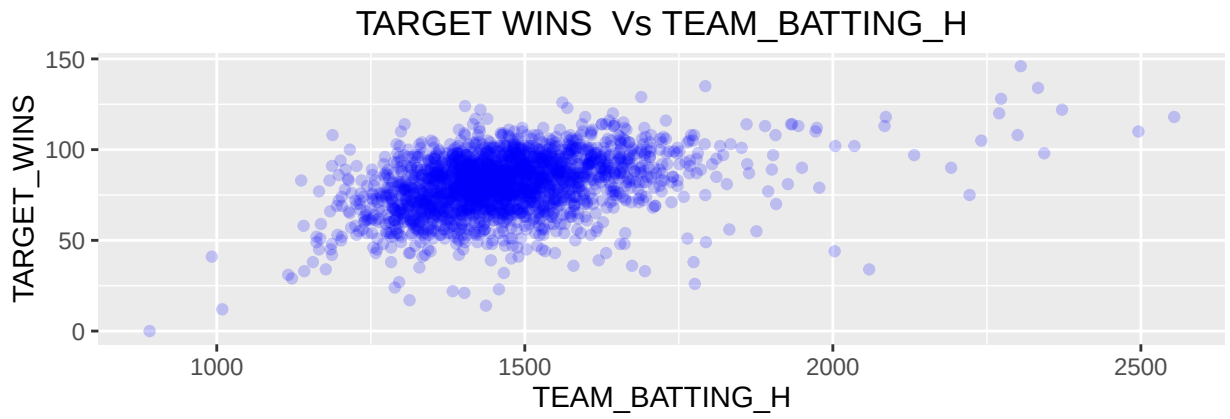
	std
TEAM_BATTING_H	144.59120
TEAM_BATTING_2B	46.80141
TEAM_BATTING_3B	27.93856
TEAM_BATTING_HR	60.54687
TEAM_BATTING_BB	122.67086
TEAM_BATTING_SO	248.52642
TEAM_BASERUN_SB	87.79117
TEAM_BASERUN_CS	22.95634
TEAM_BATTING_HBP	12.96712
TEAM_PITCHING_H	1406.84293
TEAM_PITCHING_HR	61.29875
TEAM_PITCHING_BB	166.35736
TEAM_PITCHING_SO	553.08503
TEAM_FIELDING_E	227.77097
TEAM_FIELDING_DP	26.22639

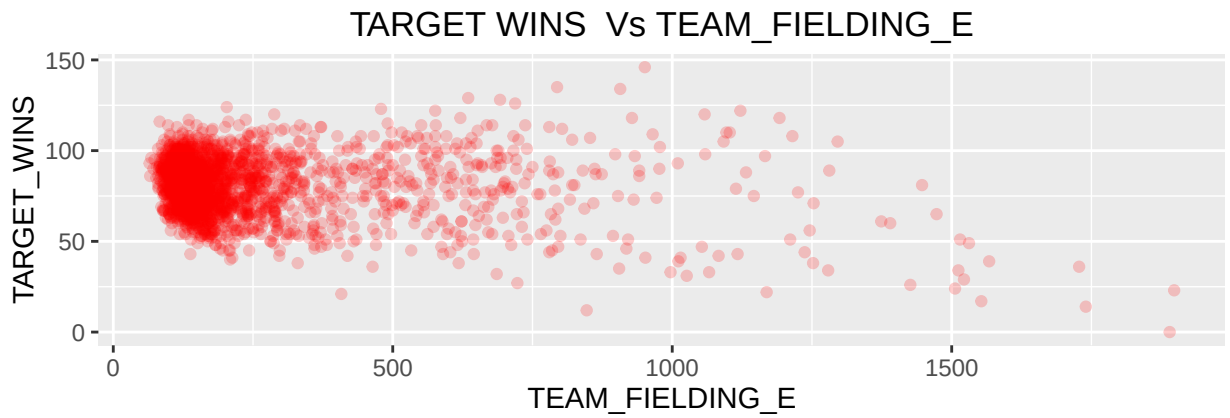
1.4 Correlation

Find correlation of Response variable with predictor variables

Variable	Correlation
TARGET_WINS	1.000
TEAM_BATTING_H	0.389
TEAM_BATTING_2B	0.289
TEAM_BATTING_3B	0.143
TEAM_BATTING_HR	0.176
TEAM_BATTING_BB	0.233
TEAM_BATTING_SO	NA
TEAM_BASERUN_SB	NA
TEAM_BASERUN_CS	NA
TEAM_BATTING_HBP	NA
TEAM_PITCHING_H	-0.110
TEAM_PITCHING_HR	0.189
TEAM_PITCHING_BB	0.124
TEAM_PITCHING_SO	NA
TEAM_FIELDING_E	-0.176
TEAM_FIELDING_DP	NA

From the above *TEAM_BATTING_H* is high positively correlated, and *TEAM_FIELDING_E* is lower side of negative corelation with the *TARGET_WINS*





2. Data Preparation

2.1 Imputation of issing data

We have noticed that there are missing values for predictors, lets impute of missing values with mean.

After imputation, the missing values should not be there.

	mb.imp
TEAM_BATTING_H	0
TEAM_BATTING_2B	0
TEAM_BATTING_3B	0
TEAM_BATTING_HR	0
TEAM_BATTING_BB	0
TEAM_BATTING_SO	0
TEAM_BASERUN_SB	0
TEAM_BASERUN_CS	0
TEAM_BATTING_HBP	0
TEAM_PITCHING_H	0
TEAM_PITCHING_HR	0
TEAM_PITCHING_BB	0
TEAM_PITCHING_SO	0
TEAM_FIELDING_E	0
TEAM_FIELDING_DP	0

Correlation of response variable to predictor variable after imputing data

Variable	Correlation
TARGET_WINS	1.000
TEAM_BATTING_H	0.389
TEAM_BATTING_2B	0.289
TEAM_BATTING_3B	0.143
TEAM_BATTING_HR	0.176
TEAM_BATTING_BB	0.233
TEAM_BATTING_SO	-0.031
TEAM_BASERUN_SB	0.123
TEAM_BASERUN_CS	0.016
TEAM_BATTING_HBP	0.016
TEAM_PITCHING_H	-0.110
TEAM_PITCHING_HR	0.189
TEAM_PITCHING_BB	0.124
TEAM_PITCHING_SO	-0.076
TEAM_FIELDING_E	-0.176
TEAM_FIELDING_DP	-0.029

3. Build Models

To fit a multiple linear regression model with TARGET_WINS as the response variable all the other predictors as the explanatory variables except 'TEAM BASERUN CS' & 'TEAM BATTING HBP' as they have very low correlation with Wins: We eliminate TEAM_BASERUN_CS

3.1 Manual elimination

```
##
## Call:
## lm(formula = 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, data = moneyballTraining)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.019  -8.640   0.148   8.354  58.658
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.095e+01  6.874e+00   3.048 0.002332 **
## TEAM_BATTING_H    4.821e-02  3.687e-03  13.075  < 2e-16 ***
## TEAM_BATTING_2B  -2.006e-02  9.152e-03  -2.192 0.028489 *
## TEAM_BATTING_3B   6.057e-02  1.676e-02   3.614 0.000308 ***
## TEAM_BATTING_HR   5.302e-02  2.743e-02   1.933 0.053347 .
## TEAM_BATTING_BB   1.037e-02  5.818e-03   1.782 0.074945 .
## TEAM_BATTING_SO  -9.408e-03  2.552e-03  -3.687 0.000232 ***
## TEAM_BASERUN_SB   2.955e-02  4.462e-03   6.623  4.4e-11 ***
## TEAM_BASERUN_CS  -1.182e-02  1.614e-02  -0.732 0.464219
## TEAM_BATTING_HBP  6.982e-02  7.303e-02   0.956 0.339166
## TEAM_PITCHING_H  -7.325e-04  3.677e-04  -1.993 0.046433 *
## TEAM_PITCHING_HR  1.483e-02  2.432e-02   0.610 0.542126
## TEAM_PITCHING_BB  7.764e-05  4.146e-03   0.019 0.985058
## TEAM_PITCHING_SO  2.846e-03  9.188e-04   3.098 0.001972 **
## TEAM_FIELDING_E  -2.118e-02  2.481e-03  -8.536  < 2e-16 ***
## TEAM_FIELDING_DP -1.208e-01  1.302e-02  -9.274  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.04 on 2260 degrees of freedom
## Multiple R-squared:  0.3192, Adjusted R-squared:  0.3147
## F-statistic: 70.65 on 15 and 2260 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO +
##     TEAM_BASERUN_SB + TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_BB +
##     TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP, data = moneyballTraining)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.906  -8.582   0.121   8.411  58.597
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       24.2467875  5.2929578   4.581 4.88e-06 ***
## TEAM_BATTING_H     0.0482145  0.0036862  13.080  < 2e-16 ***
```

```
## TEAM_BATTING_2B  -0.0203354  0.0091433  -2.224 0.026242 *
## TEAM_BATTING_3B   0.0608479  0.0167500   3.633 0.000287 ***
## TEAM_BATTING_HR   0.0549878  0.0272921   2.015 0.044045 *
## TEAM_BATTING_BB   0.0105498  0.0058145   1.814 0.069749 .
## TEAM_BATTING_SO  -0.0093113  0.0025500  -3.652 0.000267 ***
## TEAM_BASERUN_SB   0.0287308  0.0043400   6.620 4.47e-11 ***
## TEAM_PITCHING_H  -0.0007465  0.0003672  -2.033 0.042189 *
## TEAM_PITCHING_HR  0.0141572  0.0243055   0.582 0.560309
## TEAM_PITCHING_BB  0.0001870  0.0041429   0.045 0.964010
## TEAM_PITCHING_SO  0.0028358  0.0009186   3.087 0.002046 **
## TEAM_FIELDING_E  -0.0207258  0.0024210  -8.561  < 2e-16 ***
## TEAM_FIELDING_DP -0.1211711  0.0130165  -9.309  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.04 on 2262 degrees of freedom
## Multiple R-squared:  0.3188, Adjusted R-squared:  0.3149
## F-statistic: 81.42 on 13 and 2262 DF,  p-value: < 2.2e-16
```

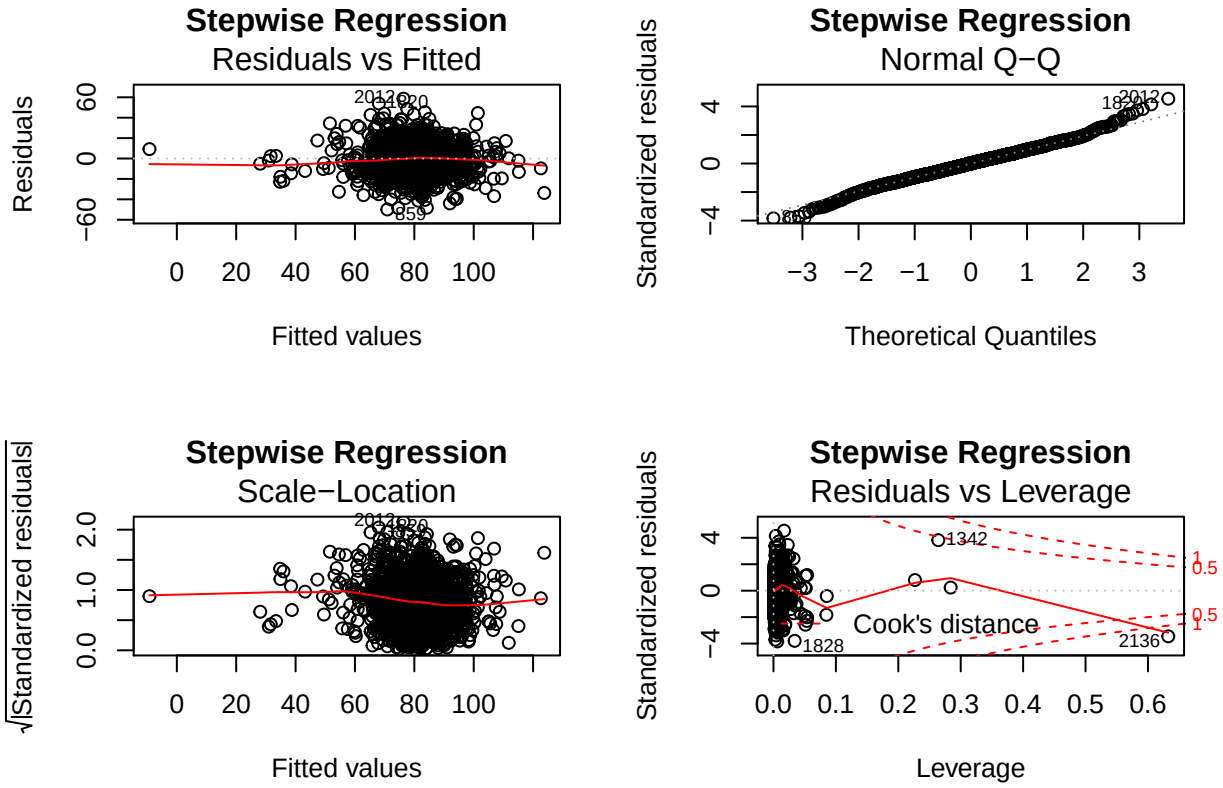
3.2 Stepwise Regression

```
## Start:  AIC=11705.6
## 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
##
##                     Df Sum of Sq    RSS   AIC
## - TEAM_PITCHING_BB   1       0.1 384298 11704
## - TEAM_PITCHING_HR   1      63.2 384361 11704
## - TEAM_BASERUN_CS    1      91.1 384389 11704
## - TEAM_BATTING_HBP   1     155.4 384454 11704
## <none>                           384298 11706
## - TEAM_BATTING_BB    1     539.7 384838 11707
## - TEAM_BATTING_HR    1     635.4 384934 11707
## - TEAM_PITCHING_H    1     675.1 384973 11708
## - TEAM_BATTING_2B    1     817.0 385115 11708
## - TEAM_PITCHING_SO   1    1632.0 385930 11713
## - TEAM_BATTING_3B    1    2220.6 386519 11717
## - TEAM_BATTING_SO    1    2312.0 386610 11717
## - TEAM_BASERUN_SB    1    7457.8 391756 11747
## - TEAM_FIELDING_E    1   12389.6 396688 11776
## - TEAM_FIELDING_DP   1   14623.6 398922 11789
## - TEAM_BATTING_H     1   29068.5 413367 11870
##
## Step:  AIC=11703.6
## 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_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq    RSS   AIC
## - TEAM_PITCHING_HR   1      87.5 384386 11702
```

```
## - TEAM_BASERUN_CS   1      91.4 384390 11702
## - TEAM_BATTING_HBP  1     155.4 384454 11702
## <none>                          384298 11704
## + TEAM_PITCHING_BB  1       0.1 384298 11706
## - TEAM_BATTING_2B   1     816.9 385115 11706
## - TEAM_BATTING_HR   1     817.1 385115 11706
## - TEAM_PITCHING_H   1     839.1 385137 11707
## - TEAM_BATTING_BB   1    1638.6 385937 11711
## - TEAM_BATTING_3B   1    2220.5 386519 11715
## - TEAM_BATTING_SO   1    2494.3 386793 11716
## - TEAM_PITCHING_SO  1    3077.2 387375 11720
## - TEAM_BASERUN_SB   1    7641.2 391939 11746
## - TEAM_FIELDING_E   1   12398.6 396697 11774
## - TEAM_FIELDING_DP  1   14627.7 398926 11787
## - TEAM_BATTING_H    1   29148.5 413447 11868
##
## Step:  AIC=11702.12
## 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_SO +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq    RSS   AIC
## - TEAM_BASERUN_CS   1      87.4 384473 11701
## - TEAM_BATTING_HBP  1     155.5 384541 11701
## <none>                          384386 11702
## + TEAM_PITCHING_HR  1      87.5 384298 11704
## + TEAM_PITCHING_BB  1      24.3 384361 11704
## - TEAM_PITCHING_H   1     764.1 385150 11705
## - TEAM_BATTING_2B   1     833.4 385219 11705
## - TEAM_BATTING_BB   1    1634.1 386020 11710
## - TEAM_BATTING_3B   1    2390.2 386776 11714
## - TEAM_BATTING_SO   1    2473.9 386860 11715
## - TEAM_PITCHING_SO  1    3134.9 387521 11719
## - TEAM_BASERUN_SB   1    7638.4 392024 11745
## - TEAM_BATTING_HR   1    8329.0 392715 11749
## - TEAM_FIELDING_E   1   12311.8 396697 11772
## - TEAM_FIELDING_DP  1   14594.6 398980 11785
## - TEAM_BATTING_H    1   29740.4 414126 11870
##
## Step:  AIC=11700.64
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_BATTING_HBP + TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP
##
##                     Df Sum of Sq    RSS   AIC
## - TEAM_BATTING_HBP  1     156.4 384629 11700
## <none>                          384473 11701
## + TEAM_BASERUN_CS   1      87.4 384386 11702
## + TEAM_PITCHING_HR  1      83.5 384390 11702
## + TEAM_PITCHING_BB  1      25.8 384447 11702
## - TEAM_PITCHING_H   1     786.3 385259 11703
## - TEAM_BATTING_2B   1     856.4 385329 11704
```

```
## - TEAM_BATTING_BB   1    1727.3 386200 11709
## - TEAM_BATTING_3B   1    2417.1 386890 11713
## - TEAM_BATTING_SO   1    2463.8 386937 11713
## - TEAM_PITCHING_SO  1    3149.8 387623 11717
## - TEAM_BASERUN_SB   1    7671.9 392145 11744
## - TEAM_BATTING_HR   1    8940.1 393413 11751
## - TEAM_FIELDING_E   1   12449.8 396923 11771
## - TEAM_FIELDING_DP  1   14628.8 399102 11784
## - TEAM_BATTING_H    1   29699.2 414172 11868
##
## Step:  AIC=11699.56
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq    RSS   AIC
## <none>                           384629 11700
## + TEAM_BATTING_HBP  1     156.4 384473 11701
## + TEAM_BASERUN_CS   1      88.3 384541 11701
## + TEAM_PITCHING_HR  1      83.5 384546 11701
## + TEAM_PITCHING_BB  1      26.2 384603 11701
## - TEAM_PITCHING_H   1     789.1 385419 11702
## - TEAM_BATTING_2B   1     856.4 385486 11703
## - TEAM_BATTING_BB   1    1748.8 386378 11708
## - TEAM_BATTING_3B   1    2410.2 387040 11712
## - TEAM_BATTING_SO   1    2434.8 387064 11712
## - TEAM_PITCHING_SO  1    3146.9 387776 11716
## - TEAM_BASERUN_SB   1    7640.6 392270 11742
## - TEAM_BATTING_HR   1    8926.2 393556 11750
## - TEAM_FIELDING_E   1   12388.9 397018 11770
## - TEAM_FIELDING_DP  1   14701.6 399331 11783
## - TEAM_BATTING_H    1   29745.8 414375 11867

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## 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
##
## Final Model:
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##
##                  Step Df     Deviance Resid. Df Resid. Dev      AIC
## 1                                         2260   384298.1 11705.60
## 2 - TEAM_PITCHING_BB  1   0.05965238      2261   384298.2 11703.60
## 3 - TEAM_PITCHING_HR  1  87.47907196      2262   384385.7 11702.12
## 4  - TEAM_BASERUN_CS  1  87.35140747      2263   384473.0 11700.64
## 5 - TEAM_BATTING_HBP  1 156.44225994      2264   384629.5 11699.56
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO +
##     TEAM_BASERUN_SB + TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP, data = moneyballTraining)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.899  -8.568   0.091   8.397  58.651
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      23.6666983  5.2220414   4.532 6.14e-06 ***
## TEAM_BATTING_H    0.0484570  0.0036621  13.232  < 2e-16 ***
## TEAM_BATTING_2B  -0.0205123  0.0091358  -2.245 0.024847 *
## TEAM_BATTING_3B   0.0624661  0.0165843   3.767 0.000170 ***
## TEAM_BATTING_HR   0.0697785  0.0096266   7.249 5.75e-13 ***
## TEAM_BATTING_BB   0.0107446  0.0033489   3.208 0.001354 **
## TEAM_BATTING_SO  -0.0093019  0.0024571  -3.786 0.000157 ***
## TEAM_BASERUN_SB   0.0287708  0.0042901   6.706 2.51e-11 ***
## TEAM_PITCHING_H  -0.0006920  0.0003211  -2.155 0.031253 *
## TEAM_PITCHING_SO  0.0028867  0.0006707   4.304 1.75e-05 ***
## TEAM_FIELDING_E  -0.0205973  0.0024120  -8.540  < 2e-16 ***
## TEAM_FIELDING_DP -0.1210083  0.0130082  -9.302  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.03 on 2264 degrees of freedom
## Multiple R-squared:  0.3186, Adjusted R-squared:  0.3153
## F-statistic: 96.25 on 11 and 2264 DF,  p-value: < 2.2e-16
```



####Model with Stepwise Regression TARGET_WINS = 23.67 + 0.048*TEAM_BATTING_H* + *-0.020*TEAM_BATTING_2B + 0.0625*TEAM_BATTING_3B* + *0.0698*TEAM_BATTING_HR + 0.011*TEAM_BATTING_BB* + *-0.009*TEAM_BATTING_SO + 0.029*TEAM_BASERUN_SB* + *-0.0007*TEAM_PITCHING_H + 0.0029*TEAM_PITCHING_SO* + *-0.0206*TEAM_FIELDING_E + -0.121*TEAM_FIELDING_DP

3.3 Backward elimination model

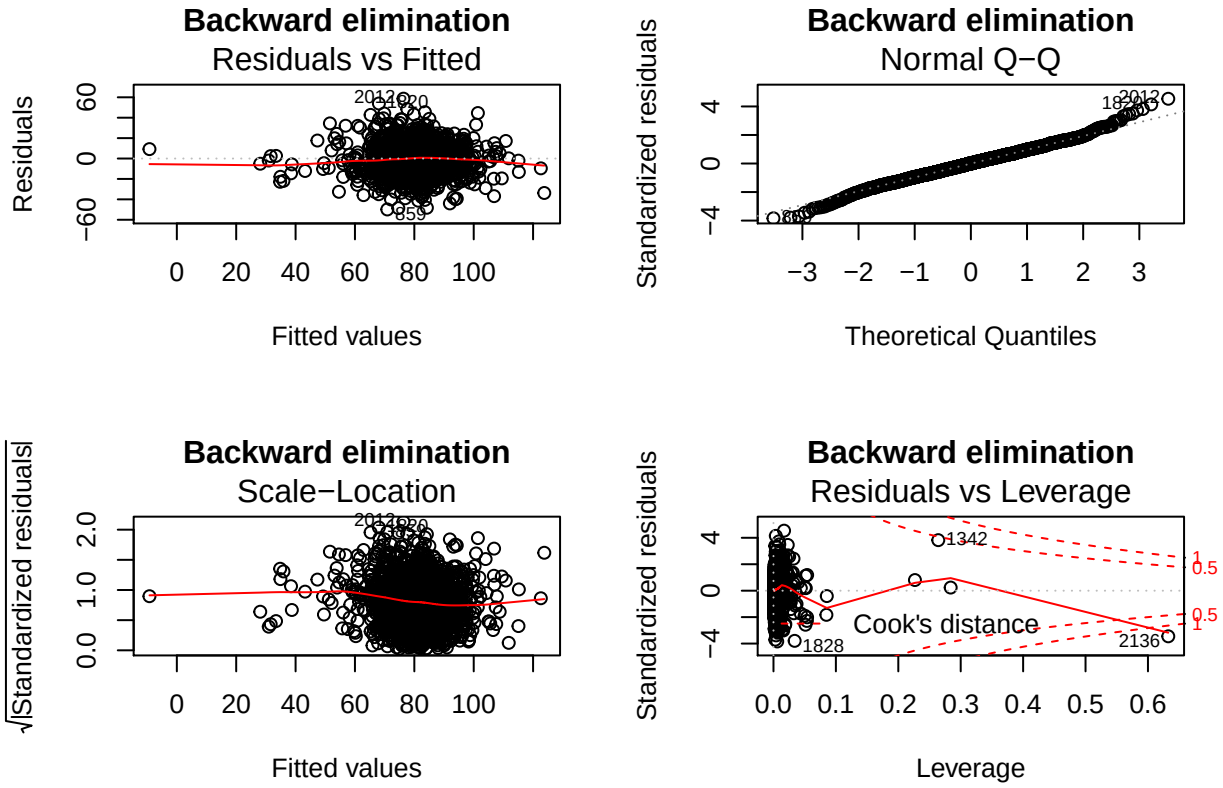
```
## Start:  AIC=11705.6
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_BASERUN_CS + TEAM_BATTING_HBP + TEAM_PITCHING_H + TEAM_PITCHING_HR +
##     TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq    RSS   AIC
## - TEAM_PITCHING_BB   1       0.1 384298 11704
## - TEAM_PITCHING_HR   1      63.2 384361 11704
## - TEAM_BASERUN_CS    1      91.1 384389 11704
## - TEAM_BATTING_HBP   1     155.4 384454 11704
## <none>                           384298 11706
## - TEAM_BATTING_BB    1     539.7 384838 11707
## - TEAM_BATTING_HR    1     635.4 384934 11707
## - TEAM_PITCHING_H    1     675.1 384973 11708
## - TEAM_BATTING_2B    1     817.0 385115 11708
## - TEAM_PITCHING_SO   1    1632.0 385930 11713
## - TEAM_BATTING_3B    1    2220.6 386519 11717
```

```
## - TEAM_BATTING_SO   1    2312.0 386610 11717
## - TEAM_BASERUN_SB   1    7457.8 391756 11747
## - TEAM_FIELDING_E   1   12389.6 396688 11776
## - TEAM_FIELDING_DP  1   14623.6 398922 11789
## - TEAM_BATTING_H    1   29068.5 413367 11870
##
## Step:  AIC=11703.6
## 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_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq    RSS   AIC
## - TEAM_PITCHING_HR  1      87.5 384386 11702
## - TEAM_BASERUN_CS   1      91.4 384390 11702
## - TEAM_BATTING_HBP  1     155.4 384454 11702
## <none>                          384298 11704
## - TEAM_BATTING_2B   1     816.9 385115 11706
## - TEAM_BATTING_HR   1     817.1 385115 11706
## - TEAM_PITCHING_H   1     839.1 385137 11707
## - TEAM_BATTING_BB   1    1638.6 385937 11711
## - TEAM_BATTING_3B   1    2220.5 386519 11715
## - TEAM_BATTING_SO   1    2494.3 386793 11716
## - TEAM_PITCHING_SO  1    3077.2 387375 11720
## - TEAM_BASERUN_SB   1    7641.2 391939 11746
## - TEAM_FIELDING_E   1   12398.6 396697 11774
## - TEAM_FIELDING_DP  1   14627.7 398926 11787
## - TEAM_BATTING_H    1   29148.5 413447 11868
##
## Step:  AIC=11702.12
## 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_SO +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq    RSS   AIC
## - TEAM_BASERUN_CS   1      87.4 384473 11701
## - TEAM_BATTING_HBP  1     155.5 384541 11701
## <none>                          384386 11702
## - TEAM_PITCHING_H   1     764.1 385150 11705
## - TEAM_BATTING_2B   1     833.4 385219 11705
## - TEAM_BATTING_BB   1    1634.1 386020 11710
## - TEAM_BATTING_3B   1    2390.2 386776 11714
## - TEAM_BATTING_SO   1    2473.9 386860 11715
## - TEAM_PITCHING_SO  1    3134.9 387521 11719
## - TEAM_BASERUN_SB   1    7638.4 392024 11745
## - TEAM_BATTING_HR   1    8329.0 392715 11749
## - TEAM_FIELDING_E   1   12311.8 396697 11772
## - TEAM_FIELDING_DP  1   14594.6 398980 11785
## - TEAM_BATTING_H    1   29740.4 414126 11870
##
## Step:  AIC=11700.64
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
```

```
##     TEAM_BATTING_HBP + TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP
##
##                     Df Sum of Sq    RSS   AIC
## - TEAM_BATTING_HBP  1     156.4 384629 11700
## <none>                          384473 11701
## - TEAM_PITCHING_H   1     786.3 385259 11703
## - TEAM_BATTING_2B   1     856.4 385329 11704
## - TEAM_BATTING_BB   1    1727.3 386200 11709
## - TEAM_BATTING_3B   1    2417.1 386890 11713
## - TEAM_BATTING_SO   1    2463.8 386937 11713
## - TEAM_PITCHING_SO  1    3149.8 387623 11717
## - TEAM_BASERUN_SB   1    7671.9 392145 11744
## - TEAM_BATTING_HR   1    8940.1 393413 11751
## - TEAM_FIELDING_E   1   12449.8 396923 11771
## - TEAM_FIELDING_DP  1   14628.8 399102 11784
## - TEAM_BATTING_H    1   29699.2 414172 11868
##
## Step:  AIC=11699.56
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq    RSS   AIC
## <none>                          384629 11700
## - TEAM_PITCHING_H   1     789.1 385419 11702
## - TEAM_BATTING_2B   1     856.4 385486 11703
## - TEAM_BATTING_BB   1    1748.8 386378 11708
## - TEAM_BATTING_3B   1    2410.2 387040 11712
## - TEAM_BATTING_SO   1    2434.8 387064 11712
## - TEAM_PITCHING_SO  1    3146.9 387776 11716
## - TEAM_BASERUN_SB   1    7640.6 392270 11742
## - TEAM_BATTING_HR   1    8926.2 393556 11750
## - TEAM_FIELDING_E   1   12388.9 397018 11770
## - TEAM_FIELDING_DP  1   14701.6 399331 11783
## - TEAM_BATTING_H    1   29745.8 414375 11867

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## 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
##
## Final Model:
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##
##                 Step Df     Deviance Resid. Df Resid. Dev      AIC
## 1                                         2260   384298.1 11705.60
```

```
## 2 - TEAM_PITCHING_BB  1   0.05965238      2261   384298.2 11703.60
## 3 - TEAM_PITCHING_HR  1  87.47907196      2262   384385.7 11702.12
## 4  - TEAM_BASERUN_CS  1  87.35140747      2263   384473.0 11700.64
## 5 - TEAM_BATTING_HBP  1 156.44225994      2264   384629.5 11699.56
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO +
##     TEAM_BASERUN_SB + TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP, data = moneyballTraining)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.899  -8.568   0.091   8.397  58.651
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      23.6666983  5.2220414   4.532 6.14e-06 ***
## TEAM_BATTING_H    0.0484570  0.0036621  13.232  < 2e-16 ***
## TEAM_BATTING_2B  -0.0205123  0.0091358  -2.245 0.024847 *
## TEAM_BATTING_3B   0.0624661  0.0165843   3.767 0.000170 ***
## TEAM_BATTING_HR   0.0697785  0.0096266   7.249 5.75e-13 ***
## TEAM_BATTING_BB   0.0107446  0.0033489   3.208 0.001354 **
## TEAM_BATTING_SO  -0.0093019  0.0024571  -3.786 0.000157 ***
## TEAM_BASERUN_SB   0.0287708  0.0042901   6.706 2.51e-11 ***
## TEAM_PITCHING_H  -0.0006920  0.0003211  -2.155 0.031253 *
## TEAM_PITCHING_SO  0.0028867  0.0006707   4.304 1.75e-05 ***
## TEAM_FIELDING_E  -0.0205973  0.0024120  -8.540  < 2e-16 ***
## TEAM_FIELDING_DP -0.1210083  0.0130082  -9.302  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.03 on 2264 degrees of freedom
## Multiple R-squared:  0.3186, Adjusted R-squared:  0.3153
## F-statistic: 96.25 on 11 and 2264 DF,  p-value: < 2.2e-16
```



###Model with Backward elimination TARGET_WINS = 23.66 + 0.048*TEAM_BATTING_H* + *-0.020*TEAM_BATTING_2B + 0.062*TEAM_BATTING_3B* + *0.0698*TEAM_BATTING_HR + 0.011*TEAM_BATTING_BB* + *-0.009*TEAM_BATTING_SO + 0.029*TEAM_BASERUN_SB* + *-0.001*TEAM_PITCHING_H + 0.002*TEAM_PITCHING_SO* + *-0.021*TEAM_FIELDING_E + -0.121*TEAM_FIELDING_DP

3.4 forward Elimination

```
## Start:  AIC=12550.76
## TARGET_WINS ~ 1
##
##                     Df Sum of Sq    RSS   AIC
## + TEAM_BATTING_H     1     85318 479178 12180
## + TEAM_BATTING_2B    1     47181 517315 12354
## + TEAM_BATTING_BB    1     30530 533966 12426
## + TEAM_PITCHING_HR   1     20167 544329 12470
## + TEAM_FIELDING_E    1     17582 546914 12481
## + TEAM_BATTING_HR    1     17516 546980 12481
## + TEAM_BATTING_3B    1     11480 553016 12506
## + TEAM_PITCHING_BB   1      8704 555792 12517
## + TEAM_BASERUN_SB    1      8536 555960 12518
## + TEAM_PITCHING_H    1      6823 557674 12525
## + TEAM_PITCHING_SO   1      3242 561254 12540
## + TEAM_BATTING_SO    1       531 563965 12551
## <none>                           564496 12551
## + TEAM_FIELDING_DP   1       470 564027 12551
```

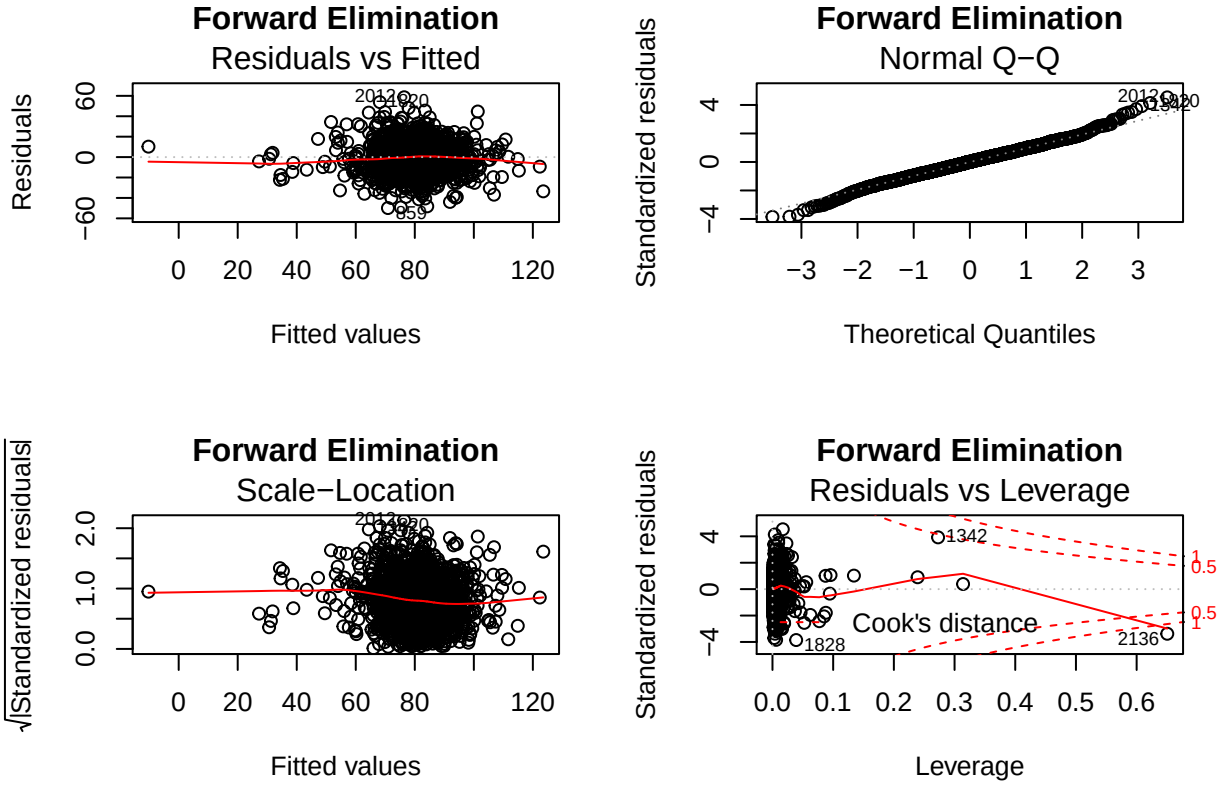
```
## + INDEX             1       250 564246 12552
## + TEAM_BATTING_HBP  1       151 564346 12552
## + TEAM_BASERUN_CS   1       137 564360 12552
##
## Step:  AIC=12179.81
## TARGET_WINS ~ TEAM_BATTING_H
##
##                     Df Sum of Sq    RSS   AIC
## + TEAM_FIELDING_E    1     47417 431762 11945
## + TEAM_BATTING_BB    1     38578 440601 11991
## + TEAM_PITCHING_H    1     32196 446983 12024
## + TEAM_BATTING_HR    1     18027 461152 12094
## + TEAM_BATTING_SO    1     14792 464387 12110
## + TEAM_PITCHING_HR   1     14654 464524 12111
## + TEAM_PITCHING_BB   1      4366 474812 12161
## + TEAM_BATTING_2B    1      4082 475097 12162
## + TEAM_BASERUN_SB    1      3537 475641 12165
## + TEAM_FIELDING_DP   1      3111 476067 12167
## <none>                           479178 12180
## + TEAM_BATTING_3B    1       387 478791 12180
## + TEAM_PITCHING_SO   1       232 478947 12181
## + TEAM_BATTING_HBP   1       184 478994 12181
## + INDEX              1       112 479066 12181
## + TEAM_BASERUN_CS    1        69 479110 12182
##
## Step:  AIC=11944.65
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_FIELDING_E
##
##                     Df Sum of Sq    RSS   AIC
## + TEAM_BASERUN_SB    1   21256.9 410505 11832
## + TEAM_FIELDING_DP   1   15847.0 415915 11862
## + TEAM_BATTING_3B    1    7944.9 423817 11904
## + TEAM_BATTING_BB    1    4858.4 426903 11921
## + TEAM_PITCHING_BB   1    3058.7 428703 11930
## + TEAM_PITCHING_H    1    2735.1 429027 11932
## + TEAM_BATTING_2B    1    2199.7 429562 11935
## + TEAM_PITCHING_SO   1     623.8 431138 11943
## <none>                           431762 11945
## + TEAM_BATTING_HBP   1     196.9 431565 11946
## + TEAM_BASERUN_CS    1     167.0 431595 11946
## + INDEX              1     134.6 431627 11946
## + TEAM_PITCHING_HR   1      35.3 431727 11946
## + TEAM_BATTING_SO    1      25.7 431736 11946
## + TEAM_BATTING_HR    1       6.5 431755 11947
##
## Step:  AIC=11831.75
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_FIELDING_E + TEAM_BASERUN_SB
##
##                     Df Sum of Sq    RSS   AIC
## + TEAM_FIELDING_DP   1    8744.5 401760 11785
## + TEAM_PITCHING_HR   1    2376.2 408129 11820
## + TEAM_BATTING_HR    1    2200.6 408304 11822
## + TEAM_BATTING_BB    1    1405.8 409099 11826
## + TEAM_PITCHING_BB   1    1287.1 409218 11827
```

```
## + TEAM_BATTING_3B   1    1021.2 409484 11828
## + TEAM_PITCHING_SO  1    1009.7 409495 11828
## + TEAM_BASERUN_CS   1     946.3 409559 11828
## + TEAM_BATTING_2B   1     506.2 409999 11831
## <none>                         410505 11832
## + INDEX             1     334.8 410170 11832
## + TEAM_BATTING_HBP  1     227.7 410277 11832
## + TEAM_PITCHING_H   1     159.9 410345 11833
## + TEAM_BATTING_SO   1      55.2 410450 11833
##
## Step:  AIC=11784.74
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_FIELDING_E + TEAM_BASERUN_SB +
##     TEAM_FIELDING_DP
##
##                     Df Sum of Sq    RSS   AIC
## + TEAM_PITCHING_HR  1    5461.8 396299 11756
## + TEAM_BATTING_HR   1    5283.5 396477 11757
## + TEAM_BATTING_BB   1    4539.2 397221 11761
## + TEAM_PITCHING_BB  1    3259.6 398501 11768
## + TEAM_BASERUN_CS   1    1459.2 400301 11778
## + TEAM_PITCHING_SO  1    1297.9 400463 11779
## <none>                         401760 11785
## + TEAM_BATTING_3B   1     343.0 401417 11785
## + INDEX             1     233.2 401527 11785
## + TEAM_BATTING_2B   1     227.2 401533 11786
## + TEAM_BATTING_HBP  1     183.3 401577 11786
## + TEAM_BATTING_SO   1     161.0 401599 11786
## + TEAM_PITCHING_H   1      47.9 401713 11786
##
## Step:  AIC=11755.59
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_FIELDING_E + TEAM_BASERUN_SB +
##     TEAM_FIELDING_DP + TEAM_PITCHING_HR
##
##                     Df Sum of Sq    RSS   AIC
## + TEAM_BATTING_3B   1    3985.6 392313 11735
## + TEAM_BATTING_BB   1    2844.2 393454 11741
## + TEAM_BATTING_SO   1    2668.8 393630 11742
## + TEAM_PITCHING_BB  1    1661.0 394638 11748
## + TEAM_BATTING_2B   1    1447.5 394851 11749
## + TEAM_PITCHING_H   1     449.1 395850 11755
## + INDEX             1     417.4 395881 11755
## + TEAM_BASERUN_CS   1     383.7 395915 11755
## <none>                         396299 11756
## + TEAM_PITCHING_SO  1     238.6 396060 11756
## + TEAM_BATTING_HBP  1     137.4 396161 11757
## + TEAM_BATTING_HR   1      39.4 396259 11757
##
## Step:  AIC=11734.58
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_FIELDING_E + TEAM_BASERUN_SB +
##     TEAM_FIELDING_DP + TEAM_PITCHING_HR + TEAM_BATTING_3B
##
##                     Df Sum of Sq    RSS   AIC
## + TEAM_BATTING_BB   1   2371.36 389942 11723
## + TEAM_BATTING_SO   1   1514.17 390799 11728
```

```
## + TEAM_PITCHING_BB  1   1358.59 390954 11729
## + TEAM_BATTING_2B   1    840.34 391473 11732
## + TEAM_BATTING_HR   1    544.04 391769 11733
## + INDEX             1    489.25 391824 11734
## + TEAM_PITCHING_SO  1    428.60 391884 11734
## <none>                          392313 11735
## + TEAM_BASERUN_CS   1    333.19 391980 11735
## + TEAM_BATTING_HBP  1    151.48 392162 11736
## + TEAM_PITCHING_H   1    137.47 392176 11736
##
## Step:  AIC=11722.78
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_FIELDING_E + TEAM_BASERUN_SB +
##     TEAM_FIELDING_DP + TEAM_PITCHING_HR + TEAM_BATTING_3B + TEAM_BATTING_BB
##
##                     Df Sum of Sq    RSS   AIC
## + TEAM_BATTING_SO   1   1049.06 388893 11719
## + TEAM_BATTING_2B   1    957.02 388985 11719
## + TEAM_PITCHING_SO  1    595.92 389346 11721
## + TEAM_BATTING_HR   1    350.73 389591 11723
## <none>                          389942 11723
## + INDEX             1    327.51 389614 11723
## + TEAM_BASERUN_CS   1    172.42 389769 11724
## + TEAM_BATTING_HBP  1    132.52 389809 11724
## + TEAM_PITCHING_BB  1    124.40 389817 11724
## + TEAM_PITCHING_H   1    105.89 389836 11724
##
## Step:  AIC=11718.65
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_FIELDING_E + TEAM_BASERUN_SB +
##     TEAM_FIELDING_DP + TEAM_PITCHING_HR + TEAM_BATTING_3B + TEAM_BATTING_BB +
##     TEAM_BATTING_SO
##
##                     Df Sum of Sq    RSS   AIC
## + TEAM_PITCHING_SO  1   1432.08 387461 11712
## + TEAM_BATTING_HR   1    714.85 388178 11716
## + TEAM_BATTING_2B   1    526.69 388366 11718
## <none>                          388893 11719
## + INDEX             1    264.84 388628 11719
## + TEAM_BASERUN_CS   1    203.00 388690 11720
## + TEAM_PITCHING_H   1    173.46 388719 11720
## + TEAM_BATTING_HBP  1    153.87 388739 11720
## + TEAM_PITCHING_BB  1     88.82 388804 11720
##
## Step:  AIC=11712.25
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_FIELDING_E + TEAM_BASERUN_SB +
##     TEAM_FIELDING_DP + TEAM_PITCHING_HR + TEAM_BATTING_3B + TEAM_BATTING_BB +
##     TEAM_BATTING_SO + TEAM_PITCHING_SO
##
##                     Df Sum of Sq    RSS   AIC
## + TEAM_PITCHING_H   1   1234.15 386226 11707
## + TEAM_BATTING_HR   1   1160.65 386300 11707
## + TEAM_BATTING_2B   1    856.73 386604 11709
## + TEAM_PITCHING_BB  1    804.87 386656 11710
## <none>                          387461 11712
## + INDEX             1    279.92 387181 11713
```

```
## + TEAM_BASERUN_CS   1    231.93 387229 11713
## + TEAM_BATTING_HBP  1    157.70 387303 11713
##
## Step:  AIC=11706.99
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_FIELDING_E + TEAM_BASERUN_SB +
##     TEAM_FIELDING_DP + TEAM_PITCHING_HR + TEAM_BATTING_3B + TEAM_BATTING_BB +
##     TEAM_BATTING_SO + TEAM_PITCHING_SO + TEAM_PITCHING_H
##
##                     Df Sum of Sq    RSS   AIC
## + TEAM_BATTING_HR   1    839.85 385387 11704
## + TEAM_BATTING_2B   1    804.52 385422 11704
## <none>                          386226 11707
## + INDEX             1    233.50 385993 11708
## + TEAM_PITCHING_BB  1    186.75 386040 11708
## + TEAM_BASERUN_CS   1    181.07 386045 11708
## + TEAM_BATTING_HBP  1    154.62 386072 11708
##
## Step:  AIC=11704.04
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_FIELDING_E + TEAM_BASERUN_SB +
##     TEAM_FIELDING_DP + TEAM_PITCHING_HR + TEAM_BATTING_3B + TEAM_BATTING_BB +
##     TEAM_BATTING_SO + TEAM_PITCHING_SO + TEAM_PITCHING_H + TEAM_BATTING_HR
##
##                     Df Sum of Sq    RSS   AIC
## + TEAM_BATTING_2B   1    840.63 384546 11701
## <none>                          385387 11704
## + INDEX             1    264.60 385122 11704
## + TEAM_BATTING_HBP  1    156.49 385230 11705
## + TEAM_BASERUN_CS   1    116.07 385271 11705
## + TEAM_PITCHING_BB  1      0.05 385387 11706
##
## Step:  AIC=11701.07
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_FIELDING_E + TEAM_BASERUN_SB +
##     TEAM_FIELDING_DP + TEAM_PITCHING_HR + TEAM_BATTING_3B + TEAM_BATTING_BB +
##     TEAM_BATTING_SO + TEAM_PITCHING_SO + TEAM_PITCHING_H + TEAM_BATTING_HR +
##     TEAM_BATTING_2B
##
##                     Df Sum of Sq    RSS   AIC
## <none>                          384546 11701
## + INDEX             1   262.936 384283 11702
## + TEAM_BATTING_HBP  1   156.401 384390 11702
## + TEAM_BASERUN_CS   1    92.336 384454 11702
## + TEAM_PITCHING_BB  1     0.346 384546 11703


##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_FIELDING_E +
##     TEAM_BASERUN_SB + TEAM_FIELDING_DP + TEAM_PITCHING_HR + TEAM_BATTING_3B +
##     TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_PITCHING_SO + TEAM_PITCHING_H +
##     TEAM_BATTING_HR + TEAM_BATTING_2B, data = moneyballTraining)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.905  -8.584   0.124   8.406  58.593
##
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      24.2348098  5.2851330   4.585 4.78e-06 ***
## TEAM_BATTING_H    0.0482055  0.0036800  13.099  < 2e-16 ***
## TEAM_FIELDING_E  -0.0207217  0.0024188  -8.567  < 2e-16 ***
## TEAM_BASERUN_SB   0.0287600  0.0042906   6.703 2.57e-11 ***
## TEAM_FIELDING_DP -0.1211603  0.0130114  -9.312  < 2e-16 ***
## TEAM_PITCHING_HR  0.0147103  0.0209846   0.701 0.483372
## TEAM_BATTING_3B   0.0608466  0.0167463   3.633 0.000286 ***
## TEAM_BATTING_BB   0.0107643  0.0033494   3.214 0.001328 **
## TEAM_BATTING_SO  -0.0093418  0.0024580  -3.800 0.000148 ***
## TEAM_PITCHING_SO  0.0028640  0.0006716   4.265 2.08e-05 ***
## TEAM_PITCHING_H  -0.0007390  0.0003281  -2.253 0.024372 *
## TEAM_BATTING_HR   0.0543985  0.0239594   2.270 0.023274 *
## TEAM_BATTING_2B  -0.0203302  0.0091405  -2.224 0.026235 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.04 on 2263 degrees of freedom
## Multiple R-squared:  0.3188, Adjusted R-squared:  0.3152
## F-statistic: 88.25 on 12 and 2263 DF,  p-value: < 2.2e-16
```



###Model with forward elimination TARGET_WINS = 24.23 + 0.048*TEAM_BATTING_H* + *-0.021*TEAM_FIELDING_E + 0.029*TEAM_BASERUN_SB* + *-0.12*TEAM_FIELDING_DP + 0.015*TEAM_PITCHING_HR* + *0.061*TEAM_BATTING_3B + 0.011*TEAM_BATTING_BB* + *-0.009*TEAM_BATTING_SO + 0.003*TEAM_PITCHING_SO* + *-0.0007*TEAM_PITCHING_H + 0.054*TEAM_BATTING_HR* + *-0.02*TEAM_BATTING_2B

Backward elimination and Stepwise Regression Model has the best Adjusted R-squared value.

So we would like select Backward elimination model

Adding the below simple regression analysis, we can remove if it does not make sense in final version.

3.5 Regression Analysis by removing high Variance Inflation Factor(VIF), and high p value predictors

```
## Observations: 2,276
## Variables: 17
## $ INDEX            <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 15, 16, 1...
## $ TARGET_WINS      <dbl> 39, 70, 86, 70, 82, 75, 80, 85, 86, 76, 78, 6...
## $ TEAM_BATTING_H   <dbl> 1445, 1339, 1377, 1387, 1297, 1279, 1244, 127...
## $ TEAM_BATTING_2B  <dbl> 194, 219, 232, 209, 186, 200, 179, 171, 197, ...
## $ TEAM_BATTING_3B  <dbl> 39, 22, 35, 38, 27, 36, 54, 37, 40, 18, 27, 3...
## $ TEAM_BATTING_HR  <dbl> 13, 190, 137, 96, 102, 92, 122, 115, 114, 96,...
## $ TEAM_BATTING_BB  <dbl> 143, 685, 602, 451, 472, 443, 525, 456, 447, ...
## $ TEAM_BATTING_SO  <dbl> 842, 1075, 917, 922, 920, 973, 1062, 1027, 92...
## $ TEAM_BASERUN_SB  <dbl> 124.7618, 37.0000, 46.0000, 43.0000, 49.0000,...
## $ TEAM_BASERUN_CS  <dbl> 52.80386, 28.00000, 27.00000, 30.00000, 39.00...
## $ TEAM_BATTING_HBP <dbl> 59.35602, 59.35602, 59.35602, 59.35602, 59.35...
## $ TEAM_PITCHING_H  <dbl> 9364, 1347, 1377, 1396, 1297, 1279, 1244, 128...
## $ TEAM_PITCHING_HR <dbl> 84, 191, 137, 97, 102, 92, 122, 116, 114, 96,...
## $ TEAM_PITCHING_BB <dbl> 927, 689, 602, 454, 472, 443, 525, 459, 447, ...
## $ TEAM_PITCHING_SO <dbl> 5456, 1082, 917, 928, 920, 973, 1062, 1033, 9...
## $ TEAM_FIELDING_E  <dbl> 1011, 193, 175, 164, 138, 123, 136, 112, 127,...
## $ TEAM_FIELDING_DP <dbl> 146.3879, 155.0000, 153.0000, 156.0000, 168.0...

## Observations: 1,707
## Variables: 17
## $ INDEX            <dbl> 695, 2, 1294, 35, 163, 2416, 218, 726, 2236, ...
## $ TARGET_WINS      <dbl> 95, 70, 88, 66, 76, 62, 67, 69, 75, 67, 81, 1...
## $ TEAM_BATTING_H   <dbl> 1615, 1339, 1355, 1468, 1354, 1367, 1489, 128...
## $ TEAM_BATTING_2B  <dbl> 325, 219, 217, 251, 226, 214, 304, 150, 258, ...
## $ TEAM_BATTING_3B  <dbl> 33, 22, 31, 23, 34, 41, 70, 76, 37, 83, 89, 6...
## $ TEAM_BATTING_HR  <dbl> 216, 190, 118, 169, 98, 130, 79, 17, 171, 22,...
## $ TEAM_BATTING_BB  <dbl> 591, 685, 611, 566, 491, 533, 523, 424, 531, ...
## $ TEAM_BATTING_SO  <dbl> 960.0000, 1075.0000, 825.0000, 1007.0000, 825...
## $ TEAM_BASERUN_SB  <dbl> 119.0000, 37.0000, 138.0000, 92.0000, 150.000...
## $ TEAM_BASERUN_CS  <dbl> 42.00000, 28.00000, 52.00000, 76.00000, 60.00...
## $ TEAM_BATTING_HBP <dbl> 53.00000, 59.35602, 59.35602, 59.35602, 59.35...
## $ TEAM_PITCHING_H  <dbl> 1615, 1347, 1355, 2068, 2069, 1375, 1566, 137...
## $ TEAM_PITCHING_HR <dbl> 216, 191, 118, 238, 150, 131, 83, 18, 192, 26...
## $ TEAM_PITCHING_BB <dbl> 591, 689, 611, 797, 750, 536, 550, 455, 597, ...
## $ TEAM_PITCHING_SO <dbl> 960.0000, 1082.0000, 825.0000, 1419.0000, 126...
## $ TEAM_FIELDING_E  <dbl> 133, 193, 127, 107, 156, 120, 199, 279, 119, ...
## $ TEAM_FIELDING_DP <dbl> 190.0000, 155.0000, 106.0000, 155.0000, 142.0...

## Observations: 409
## Variables: 17
```

```
## $ INDEX            <dbl> 1294, 307, 1107, 2294, 40, 309, 634, 615, 813...
## $ TARGET_WINS      <dbl> 88, 67, 107, 92, 70, 71, 89, 101, 73, 106, 81...
## $ TEAM_BATTING_H   <dbl> 1355, 1263, 1725, 1731, 1404, 1325, 1526, 165...
## $ TEAM_BATTING_2B  <dbl> 217, 192, 194, 275, 248, 162, 231, 277, 261, ...
## $ TEAM_BATTING_3B  <dbl> 31, 83, 67, 93, 22, 76, 59, 103, 28, 83, 28, ...
## $ TEAM_BATTING_HR  <dbl> 118, 22, 4, 88, 158, 10, 67, 39, 137, 140, 13...
## $ TEAM_BATTING_BB  <dbl> 611, 352, 79, 404, 511, 596, 609, 495, 485, 5...
## $ TEAM_BATTING_SO  <dbl> 825.0000, 566.0000, 0.0000, 479.0000, 1022.00...
## $ TEAM_BASERUN_SB  <dbl> 138.0000, 385.0000, 124.7618, 100.0000, 71.00...
## $ TEAM_BASERUN_CS  <dbl> 52.00000, 52.80386, 52.80386, 100.00000, 45.0...
## $ TEAM_BATTING_HBP <dbl> 59.35602, 59.35602, 59.35602, 59.35602, 59.35...
## $ TEAM_PITCHING_H  <dbl> 1355, 1493, 3408, 1833, 1404, 6313, 1605, 174...
## $ TEAM_PITCHING_HR <dbl> 118, 26, 8, 93, 158, 48, 70, 41, 137, 147, 13...
## $ TEAM_PITCHING_BB <dbl> 611, 416, 156, 428, 511, 2840, 641, 521, 485,...
## $ TEAM_PITCHING_SO <dbl> 825.0000, 669.0000, 0.0000, 507.0000, 1022.00...
## $ TEAM_FIELDING_E  <dbl> 127, 545, 853, 232, 106, 519, 176, 204, 120, ...
## $ TEAM_FIELDING_DP <dbl> 106.0000, 146.3879, 146.3879, 138.0000, 156.0...
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ . - INDEX, data = moneyballTraining)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.256  -8.414   0.123   8.236  59.102
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      29.6372335  7.8670546   3.767 0.000171 ***
## TEAM_BATTING_H    0.0415192  0.0041960   9.895  < 2e-16 ***
## TEAM_BATTING_2B  -0.0089020  0.0104744  -0.850 0.395510
## TEAM_BATTING_3B   0.0610804  0.0190065   3.214 0.001335 **
## TEAM_BATTING_HR   0.0320857  0.0327473   0.980 0.327326
## TEAM_BATTING_BB   0.0115603  0.0067290   1.718 0.085983 .
## TEAM_BATTING_SO  -0.0124255  0.0028883  -4.302 1.79e-05 ***
## TEAM_BASERUN_SB   0.0325476  0.0050852   6.400 2.00e-10 ***
## TEAM_BASERUN_CS  -0.0119215  0.0183313  -0.650 0.515563
## TEAM_BATTING_HBP  0.0679174  0.0830422   0.818 0.413549
## TEAM_PITCHING_H  -0.0004515  0.0004138  -1.091 0.275372
## TEAM_PITCHING_HR  0.0454008  0.0291686   1.556 0.119777
## TEAM_PITCHING_BB -0.0039184  0.0048527  -0.807 0.419513
## TEAM_PITCHING_SO  0.0017320  0.0009476   1.828 0.067767 .
## TEAM_FIELDING_E  -0.0210516  0.0028492  -7.389 2.32e-13 ***
## TEAM_FIELDING_DP -0.1117512  0.0146451  -7.631 3.88e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.94 on 1691 degrees of freedom
## Multiple R-squared:  0.3235, Adjusted R-squared:  0.3175
## F-statistic: 53.91 on 15 and 1691 DF,  p-value: < 2.2e-16
```

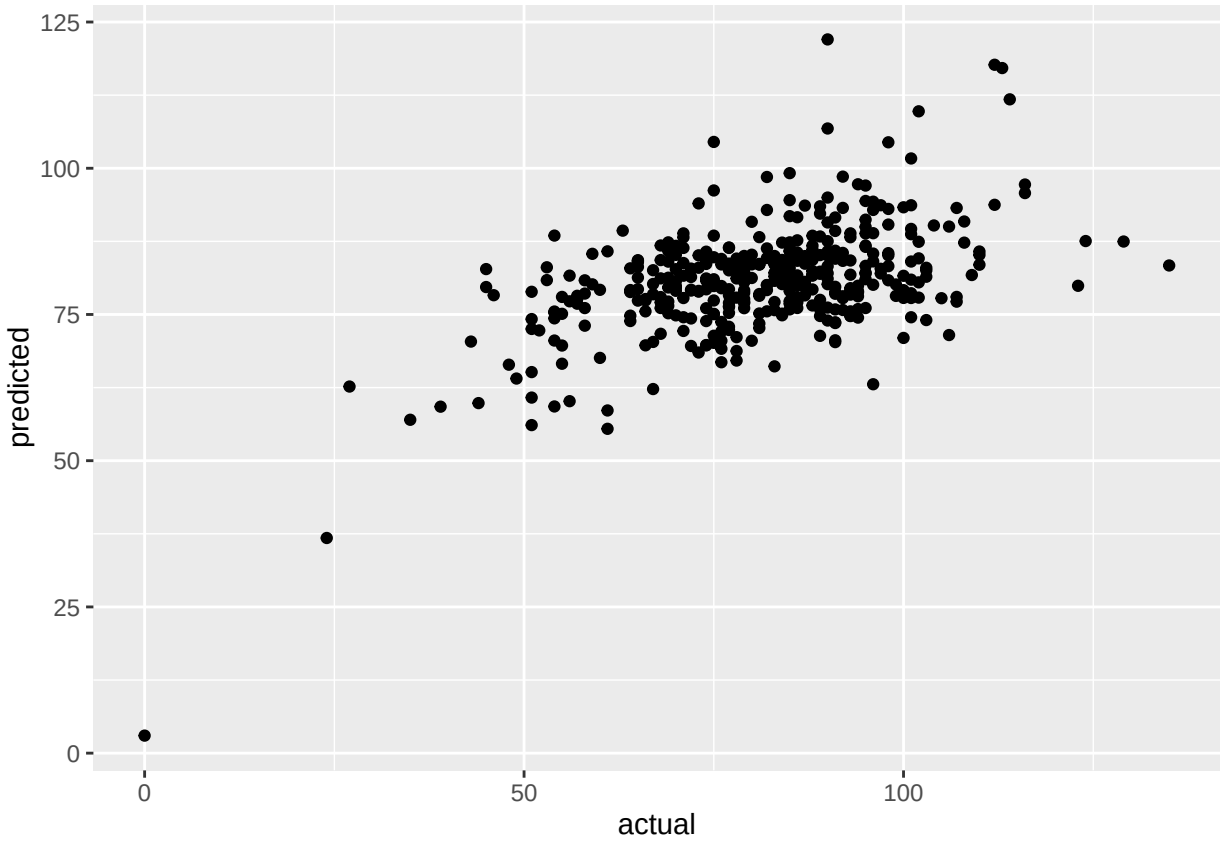
```
##   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B  TEAM_BATTING_HR
##         3.873960         2.463035         2.956973        40.345137
##  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS
```

```
##         6.930289         5.056772         1.932703         1.205459
## TEAM_BATTING_HBP  TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
##         1.002919         3.609783        32.787751         6.330015
## TEAM_PITCHING_SO  TEAM_FIELDING_E TEAM_FIELDING_DP
##         2.627914         4.350256         1.374783

##  TEAM_BATTING_HR TEAM_PITCHING_HR  TEAM_BATTING_BB TEAM_PITCHING_BB
##        40.345137        32.787751         6.930289         6.330015
##  TEAM_BATTING_SO  TEAM_FIELDING_E   TEAM_BATTING_H  TEAM_PITCHING_H
##         5.056772         4.350256         3.873960         3.609783
##  TEAM_BATTING_3B TEAM_PITCHING_SO  TEAM_BATTING_2B  TEAM_BASERUN_SB
##         2.956973         2.627914         2.463035         1.932703
## TEAM_FIELDING_DP  TEAM_BASERUN_CS TEAM_BATTING_HBP
##         1.374783         1.205459         1.002919

##
## Call:
## lm(formula = TARGET_WINS ~ . - INDEX - TEAM_BATTING_HR - TEAM_PITCHING_HR -
##     TEAM_BATTING_BB - TEAM_PITCHING_BB, data = moneyballTraining)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.362  -8.338   0.194   8.548  52.495
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      14.1463875  7.2618689   1.948   0.0516 .
## TEAM_BATTING_H    0.0541570  0.0038717  13.988  < 2e-16 ***
## TEAM_BATTING_2B  -0.0065629  0.0106287  -0.617   0.5370
## TEAM_BATTING_3B   0.0225740  0.0182476   1.237   0.2162
## TEAM_BATTING_SO  -0.0004309  0.0023307  -0.185   0.8533
## TEAM_BASERUN_SB   0.0298441  0.0047388   6.298 3.84e-10 ***
## TEAM_BASERUN_CS  -0.0440496  0.0181820  -2.423   0.0155 *
## TEAM_BATTING_HBP  0.0707726  0.0845774   0.837   0.4028
## TEAM_PITCHING_H  -0.0002504  0.0003609  -0.694   0.4878
## TEAM_PITCHING_SO  0.0006522  0.0007681   0.849   0.3959
## TEAM_FIELDING_E  -0.0276826  0.0024992 -11.077  < 2e-16 ***
## TEAM_FIELDING_DP -0.0751954  0.0140693  -5.345 1.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.19 on 1695 degrees of freedom
## Multiple R-squared:  0.2962, Adjusted R-squared:  0.2916
## F-statistic: 64.84 on 11 and 1695 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = TARGET_WINS ~ . - INDEX - TEAM_BATTING_HR - TEAM_PITCHING_HR -
##     TEAM_BATTING_BB - TEAM_PITCHING_BB - TEAM_BATTING_3B, data = moneyballTraining)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.598  -8.401   0.246   8.467  52.972
```

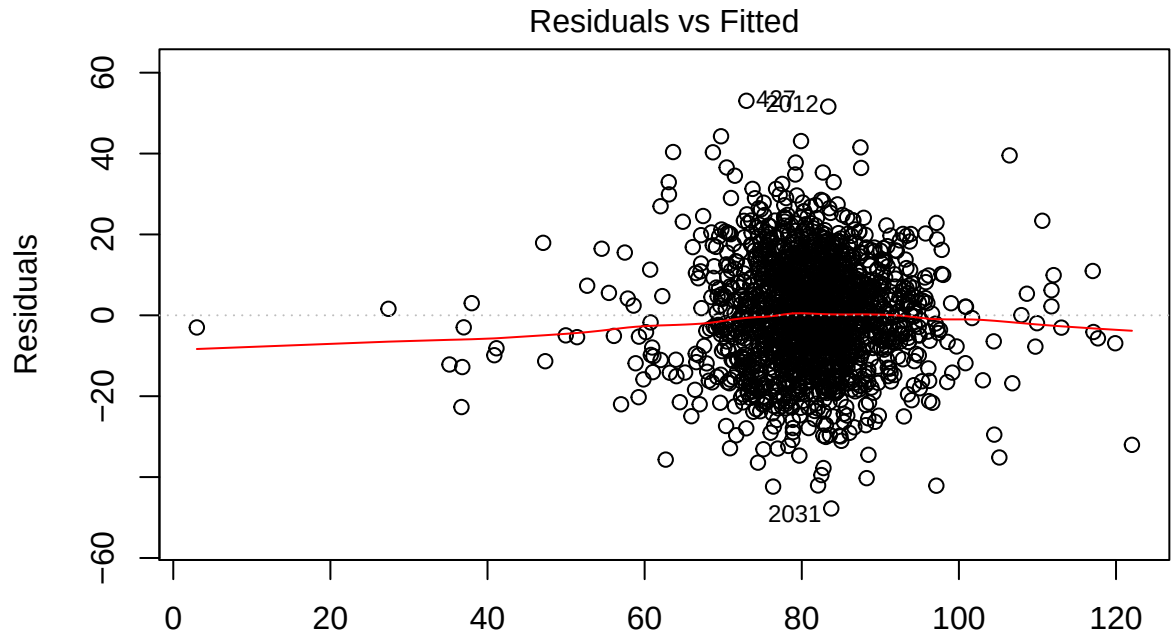
```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      14.6756446  7.2503893   2.024   0.0431 *
## TEAM_BATTING_H    0.0555729  0.0036992  15.023  < 2e-16 ***
## TEAM_BATTING_2B  -0.0080499  0.0105621  -0.762   0.4461
## TEAM_BATTING_SO  -0.0016383  0.0021169  -0.774   0.4391
## TEAM_BASERUN_SB   0.0317127  0.0044923   7.059 2.43e-12 ***
## TEAM_BASERUN_CS  -0.0432609  0.0181736  -2.380   0.0174 *
## TEAM_BATTING_HBP  0.0696636  0.0845858   0.824   0.4103
## TEAM_PITCHING_H  -0.0003474  0.0003524  -0.986   0.3243
## TEAM_PITCHING_SO  0.0007803  0.0007613   1.025   0.3055
## TEAM_FIELDING_E  -0.0272549  0.0024755 -11.010  < 2e-16 ***
## TEAM_FIELDING_DP -0.0776398  0.0139320  -5.573 2.91e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.19 on 1696 degrees of freedom
## Multiple R-squared:  0.2955, Adjusted R-squared:  0.2914
## F-statistic: 71.14 on 10 and 1696 DF,  p-value: < 2.2e-16


##
## Call:
## lm(formula = TARGET_WINS ~ . - INDEX - TEAM_BATTING_HR - TEAM_PITCHING_HR -
##     TEAM_BATTING_BB - TEAM_PITCHING_BB - TEAM_BATTING_3B - TEAM_BATTING_SO -
##     TEAM_BATTING_HBP - TEAM_PITCHING_H - TEAM_BATTING_2B, data = moneyballTraining)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.756  -8.522   0.164   8.447  53.047
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      18.4247488  3.9729984   4.637 3.80e-06 ***
## TEAM_BATTING_H    0.0536612  0.0023834  22.514  < 2e-16 ***
## TEAM_BASERUN_SB   0.0331186  0.0042952   7.711 2.12e-14 ***
## TEAM_BASERUN_CS  -0.0412880  0.0178893  -2.308   0.0211 *
## TEAM_PITCHING_SO  0.0001049  0.0006167   0.170   0.8649
## TEAM_FIELDING_E  -0.0272878  0.0015775 -17.298  < 2e-16 ***
## TEAM_FIELDING_DP -0.0796730  0.0138551  -5.750 1.05e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.19 on 1700 degrees of freedom
## Multiple R-squared:  0.2942, Adjusted R-squared:  0.2917
## F-statistic: 118.1 on 6 and 1700 DF,  p-value: < 2.2e-16
```

3.5.1. RMSE - Root Mean Sqared Error (verification with test data)

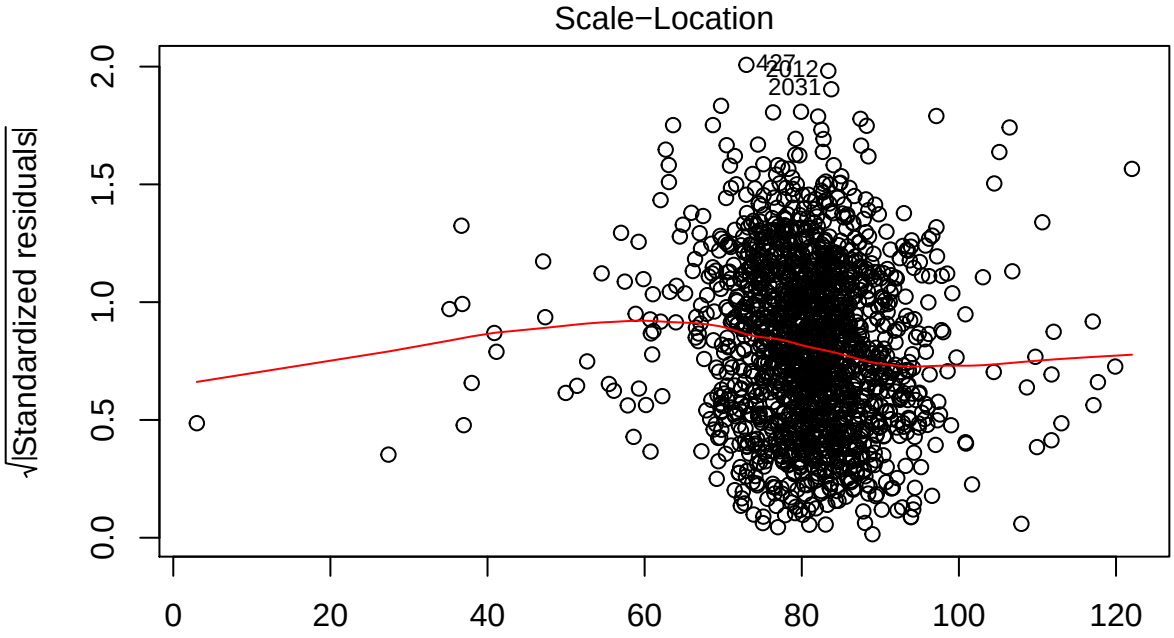
```
## [1] 13.9526
```


125
100
75
50
25
0
predicted
0
50
100
actual

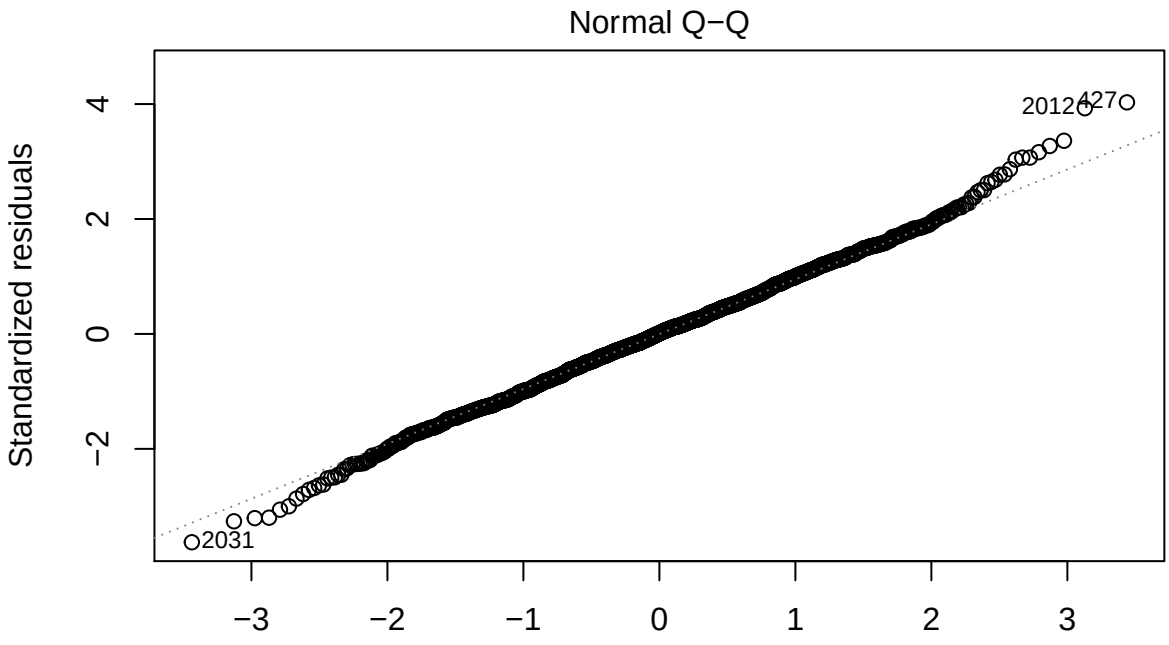
3.5.2 Diagnostic plots, check for linearity, normality is justified for residuals ...



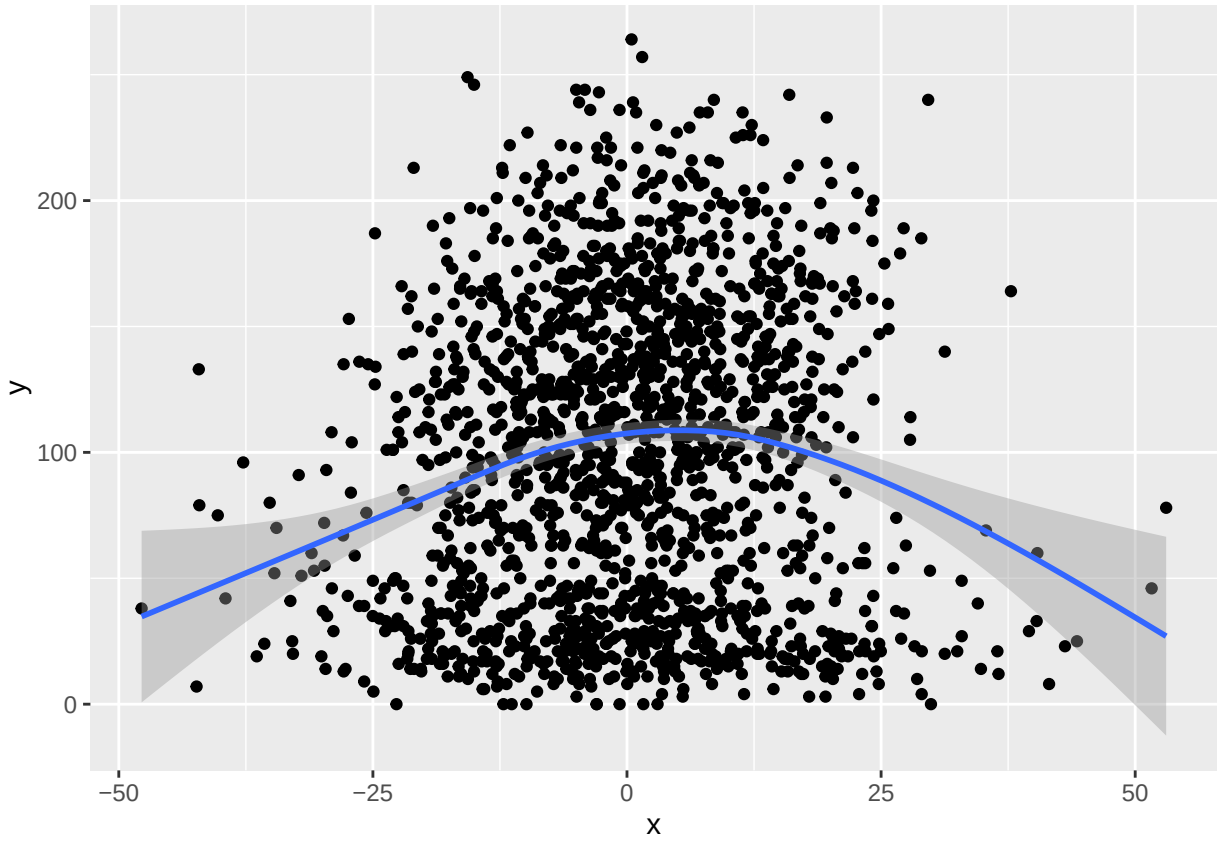
TARGET_WINS ~ . − INDEX − TEAM_BATTING_HR − TEAM_PITCHING_HR − TEAM_



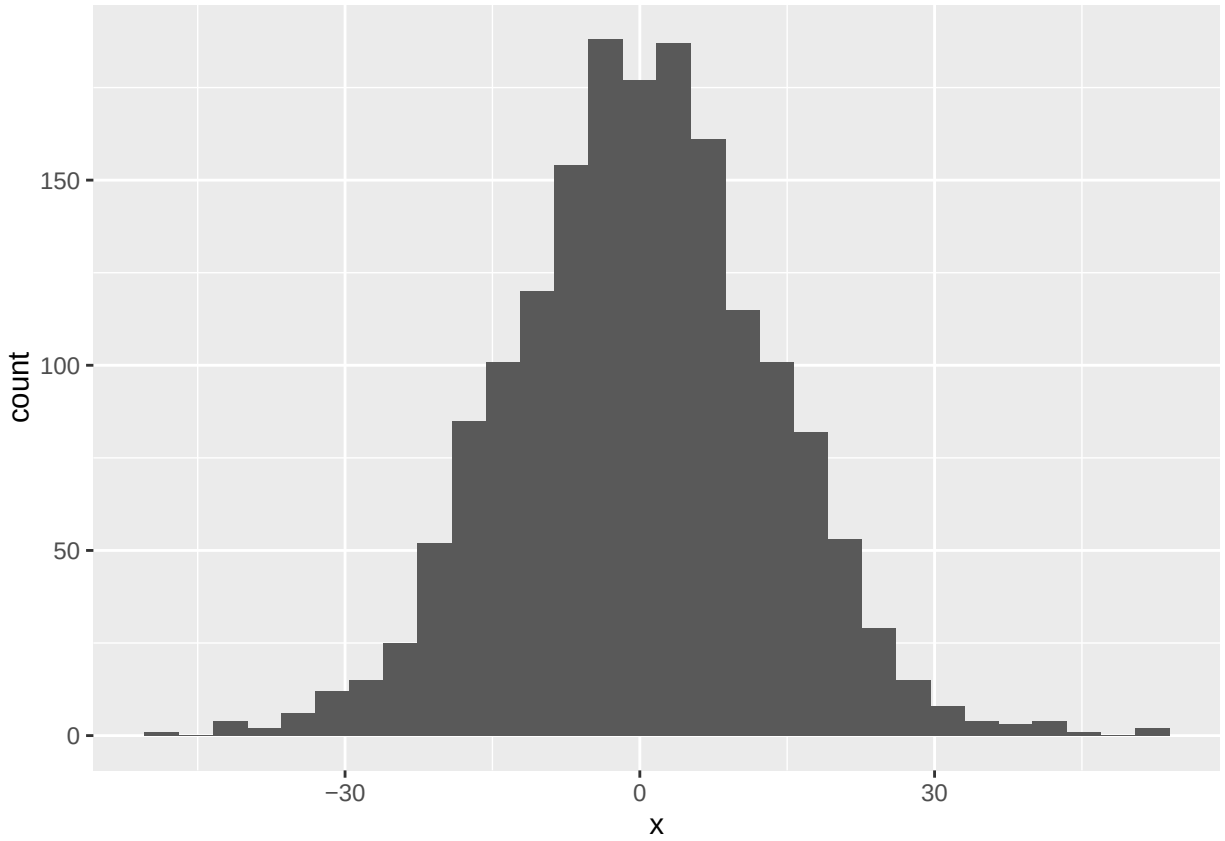
TARGET_WINS ~ . − INDEX − TEAM_BATTING_HR − TEAM_PITCHING_HR − TEAM_

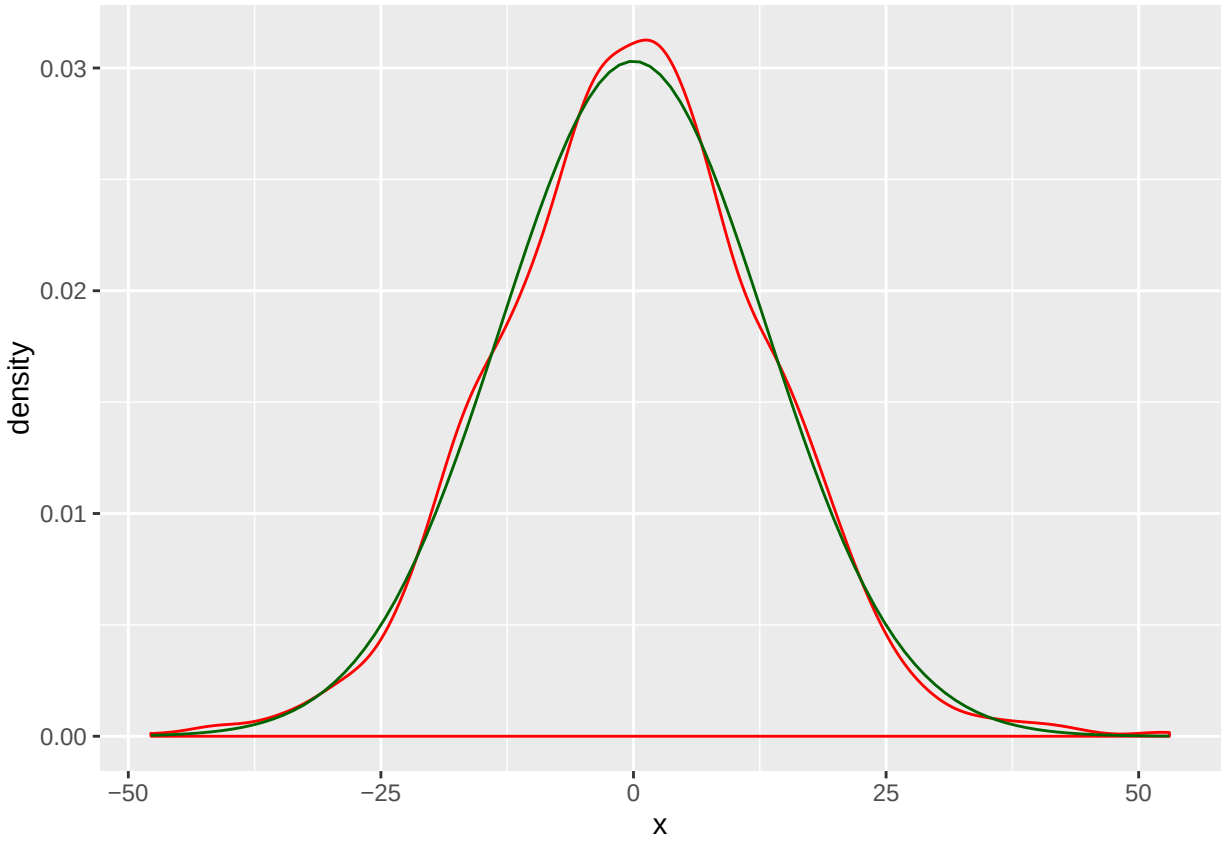


TARGET_WINS ~ . − INDEX − TEAM_BATTING_HR − TEAM_PITCHING_HR − TEAM_




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


150
100
50
0
count
−30
0
30
x


0.03
0.02
0.01
0.00
density
−50
−25
0
25
50
x

4. Selection

```
##            INDEX   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               18               13
##  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H TEAM_PITCHING_HR
##               87              240                0                0
## TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E TEAM_FIELDING_DP
##                0               18                0               31
```

```
##            INDEX   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                0                0
##  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H TEAM_PITCHING_HR
##                0                0                0                0
## TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E TEAM_FIELDING_DP
##                0                0                0                0
```

```
##        1        2        3        4        5        6
## 67.15418 67.47054 76.93439 88.00010 70.85300 73.25362
```

A. Appendix

```
library(RCurl)
library(dplyr)
library(ggplot2)
library(gridExtra)
library(gridExtra)
library(psych)
library(reshape)
library(MASS)
library(car)
library(recommenderlab)
library(knitr)
# opts_chunk$set(tidy.opts=list(width.cutoff=80),tidy=TRUE)

moneyballTraining <- read.csv("https://raw.githubusercontent.com/Nguyver/DATA621-HW/master/HW1/moneyball
    header = TRUE, sep = ",", stringsAsFactors = FALSE)

summary(moneyballTraining[3:17])

moneyball.NA <- apply(moneyballTraining[3:17], 2, function(x) sum(is.na(x)))
moneyball.missing <- cbind(moneyball.NA, moneyball.NA/nrow(moneyballTraining))
colnames(moneyball.missing) <- c("Missing", "Percentage")
kable(moneyball.missing)

# Explore independent variable TEAM_BATTING_H
g_tbh <- ggplot(data = moneyballTraining) + geom_histogram(aes(x = TEAM_BATTING_H),
```

```
    binwidth = 0.5) + theme(axis.text = element_text(size = 8),
    axis.title = element_text(size = 8))

g_b2b <- ggplot(data = moneyballTraining) + geom_histogram(aes(x = TEAM_BATTING_2B),
    binwidth = 0.5) + theme(axis.text = element_text(size = 8),
    axis.title = element_text(size = 8))

g_brsb <- ggplot(data = moneyballTraining) + geom_histogram(aes(x = TEAM_BASERUN_SB),
    binwidth = 0.5) + theme(axis.text = element_text(size = 8),
    axis.title = element_text(size = 8))

g_tph <- ggplot(data = moneyballTraining) + geom_histogram(aes(x = TEAM_PITCHING_H),
    binwidth = 0.5) + theme(axis.text = element_text(size = 8),
    axis.title = element_text(size = 8))

g_tps <- ggplot(data = moneyballTraining) + geom_histogram(aes(x = TEAM_PITCHING_SO),
    binwidth = 0.5) + theme(axis.text = element_text(size = 8),
    axis.title = element_text(size = 8))

g_tfe <- ggplot(data = moneyballTraining) + geom_histogram(aes(x = TEAM_FIELDING_E),
    binwidth = 0.5) + theme(axis.text = element_text(size = 8),
    axis.title = element_text(size = 8))

g_tfd <- ggplot(data = moneyballTraining) + geom_histogram(aes(x = TEAM_FIELDING_DP),
    binwidth = 0.5) + theme(axis.text = element_text(size = 8),
    axis.title = element_text(size = 8))

g_tbhr <- ggplot(data = moneyballTraining) + geom_histogram(aes(x = TEAM_BATTING_HR),
    binwidth = 0.5) + theme(axis.text = element_text(size = 8),
    axis.title = element_text(size = 8))

g_tphLg <- ggplot(data = moneyballTraining) + geom_histogram(aes(x = log(TEAM_PITCHING_H)),
    binwidth = 0.5) + theme(axis.text = element_text(size = 8),
    axis.title = element_text(size = 8))

g_tpsLg <- ggplot(data = moneyballTraining) + geom_histogram(aes(x = log(TEAM_PITCHING_SO)),
    binwidth = 0.5) + theme(axis.text = element_text(size = 8),
    axis.title = element_text(size = 8))

grid.arrange(g_tbh, g_b2b, g_brsb, g_tph, g_tps, g_tfe, g_tfd,
    g_tbhr, g_tphLg, g_tpsLg, ncol = 2)

meltMoneyBallTraining <- melt(moneyballTraining[3:17])
ggplot(meltMoneyBallTraining, aes(factor(variable), value)) +
    geom_boxplot() + facet_wrap(~variable, scale = "free") +
    theme(axis.text = element_text(size = 8), axis.title = element_text(size = 8))

getStandardDev <- function(moneyballTraining) {
    stdDevs <- SD(moneyballTraining[3:17])
    par(mai = c(3, 1.2, 1, 1))

    # transformed the y, due to high variances.
    barplot(stdDevs[order(stdDevs, decreasing = T)], log = "y",
```

```
        las = 2, main = "Std Dev of Predictors", xlab = "", ylab = "Log(SD)",
        cex.axis = 0.8, cex.names = 0.8)

    return(stdDevs)
}

std <- getStandardDev(moneyballTraining)
kable(as.data.frame(std))

corData <- round(cor(moneyballTraining), 3)  # rounding makes it easier to look at
t.corData <- t(corData[2, c(2:17)])  # we are only interested on correlation of Team win against all ot
moneyballTraining.cor <- melt(t.corData)  # convert the wide format to long form for ease of read
moneyballTraining.cor <- moneyballTraining.cor[, 2:3]
colnames(moneyballTraining.cor) <- c("Variable", "Correlation")

kable(moneyballTraining.cor)

g1 = ggplot(data = moneyballTraining) + geom_point(aes(x = TEAM_BATTING_H,
    y = TARGET_WINS), alpha = 0.2, color = "blue") + ggtitle("TARGET WINS  Vs TEAM_BATTING_H")

g2 = ggplot(data = moneyballTraining) + geom_point(aes(x = TEAM_FIELDING_E,
    y = TARGET_WINS), alpha = 0.2, color = "red") + ggtitle("TARGET WINS  Vs TEAM_FIELDING_E")

grid.arrange(g1, g2, nrow = 2)
# similarly other specific independent variables Vs target
# wins correlation diagram

# Replacing Missing Values In dataset with column mean
for (i in 1:ncol(moneyballTraining)) {
    moneyballTraining[is.na(moneyballTraining[, i]), i] <- mean(moneyballTraining[,
        i], na.rm = TRUE)
}

mb.imp <- apply(moneyballTraining[3:17], 2, function(x) sum(is.na(x)))
# colnames(mb.imp) <- c('# Missing')
kable(as.data.frame(mb.imp))

corData.imp <- round(cor(moneyballTraining), 3)  # rounding makes it easier to look at
t.corData.imp <- t(corData.imp[2, c(2:17)])  # we are only interested on correlation of Team win agains
moneyballTraining.cor.imp <- melt(t.corData.imp)  # convert the wide format to long form for ease of re
moneyballTraining.cor.imp <- moneyballTraining.cor.imp[, 2:3]

colnames(moneyballTraining.cor.imp) <- c("Variable", "Correlation")
kable(moneyballTraining.cor.imp)
```