Moneyball

Jorge Fernandes

MSDS 411

24 April 2018

library(e1071) # to understand skewness  
library(dplyr)  
library(stringr) # Used to rename the columns by removing the word team from the column header  
library(VIM) # To understand NAs  
library(caret)

## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'zone/tz/2018c.  
## 1.0/zoneinfo/America/New\_York'

library(mice) # Imputation   
library(missForest) # Imputation   
library(MASS) # to use for robust Linear Regression.

# browse to the data  
moneyball = read.csv('/Users/legs\_jorge/Documents/Data Science Projects/MSDS\_Northwestern/MSDS 411/Unit 01 Moneyball Baseball Problem/Data/moneyball.csv', header = T)  
colnames(moneyball) <- str\_replace\_all(colnames(moneyball),"TEAM\_","") %>%   
 tolower() # Fixing column names

## Introduction

The moneyball dataset has sparked many companies, teams, and organizations to understand and utilize the data they generate/gather. This project highlights many pitfalls that those same individuals fall into simply because they forgot to do the due diligence and prepare the data before modeling.  
This paper will focus on;  
1. Data Exploration  
2. Data Transformation  
3. Model Building  
4. How to select the best model

## Data Exploration

### Step 1: Are there lots of NAs in the data?

R gives us a lot of ways to understand the distribution of Nulls within the data. Let's first try to calculate the percentage of Null values to the total number of observation.

NAPerc <-  
 sapply(moneyball, function(x)  
 (sum(is.na(x)) / length(x)) \* 100) %>%  
 data.frame()  
NAPerc$Column <- rownames(NAPerc)  
colnames(NAPerc) <- c("NA\_Perc", "Col\_Name")  
  
# Trying to understand the percentage of NAs per Column  
NA\_col <- subset(NAPerc, NA\_Perc > 0) %>% arrange(desc(NA\_Perc))  
NA\_col

## NA\_Perc Col\_Name  
## 1 91.608084 batting\_hbp  
## 2 33.919156 baserun\_cs  
## 3 12.565905 fielding\_dp  
## 4 5.755712 baserun\_sb  
## 5 4.481547 batting\_so  
## 6 4.481547 pitching\_so

Let's look at the pattern of missing data to try to get more insights. It's clear that batting\_hbp is going to be a problematic column with 92% of the data missing. Before we start the imputation or deleting variables, let's try to understand why we have missing data.

Let's use the mice package to help us understant how all the NAs behave in the data. mice provides a handy function called md.pattern that allows one to understand the pattern of missing data. Hopefully by looking at the pattern, we can have an idea on why the data could be missing.

md.pattern(moneyball) %>% data.frame()

## index target\_wins batting\_h batting\_2b batting\_3b batting\_hr  
## 191 1 1 1 1 1 1  
## 1295 1 1 1 1 1 1  
## 349 1 1 1 1 1 1  
## 18 1 1 1 1 1 1  
## 53 1 1 1 1 1 1  
## 190 1 1 1 1 1 1  
## 102 1 1 1 1 1 1  
## 78 1 1 1 1 1 1  
## 0 0 0 0 0 0  
## batting\_bb pitching\_h pitching\_hr pitching\_bb fielding\_e batting\_so  
## 191 1 1 1 1 1 1  
## 1295 1 1 1 1 1 1  
## 349 1 1 1 1 1 1  
## 18 1 1 1 1 1 1  
## 53 1 1 1 1 1 1  
## 190 1 1 1 1 1 1  
## 102 1 1 1 1 1 0  
## 78 1 1 1 1 1 1  
## 0 0 0 0 0 102  
## pitching\_so baserun\_sb fielding\_dp baserun\_cs batting\_hbp V18  
## 191 1 1 1 1 1 0  
## 1295 1 1 1 1 0 1  
## 349 1 1 1 0 0 2  
## 18 1 1 0 1 0 2  
## 53 1 0 1 0 0 3  
## 190 1 1 0 0 0 3  
## 102 0 1 1 0 0 4  
## 78 1 0 0 0 0 4  
## 102 131 286 772 2085 3478

The **first column** of the output shows the number of unique missing data patterns. There are 191 observations with nonmissing values, and there are 1295 observations with nonmissing values except for the variable batting\_hbp. The **rightmost column** shows the number of *missing variables* in a particular missing pattern. For example, the first row has no missing value and it is “0” in the row. The **last row** counts the number of missing values for each variable. For example, the variable pitching\_bb contains no missing values and the variable batting\_so contains 102 missing values. This table can be helpful when you decide to drop some observations with missing variables exceeding a preset threshold.

After careful analysis, the decision is to keep batting\_hbp. Because I want to transform it into a binary variable, and will keep it out until all the imputation is done.

batting\_hbp\_bi <- if\_else(is.na(moneyball$batting\_hbp),0,1)  
batting\_hbp <- moneyball$batting\_hbp  
moneyball\_trans <- subset(moneyball, select = -c(batting\_hbp))

Let's impute and treat the data for missing values before testing it for multicollinearity.

The missForest package will be the package used to help us with this task. missForest is an implementation of random forest algorithm. It’s a non parametric imputation method applicable to various variable types. A great resource to understand this techinique is found [here](https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/).

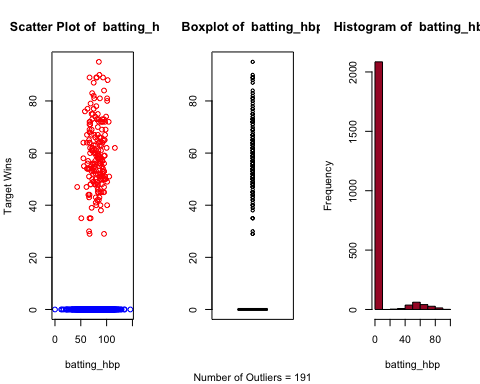
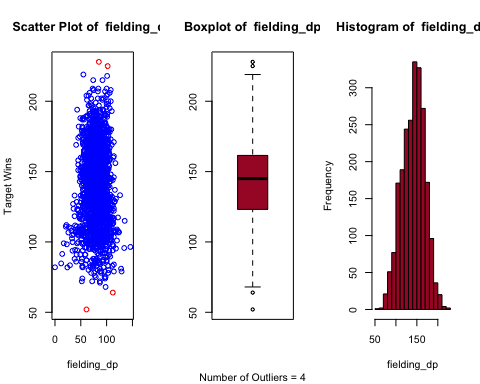
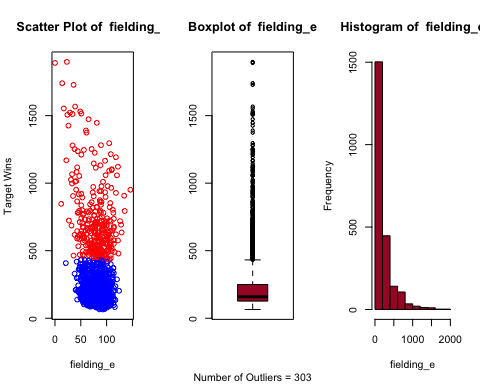
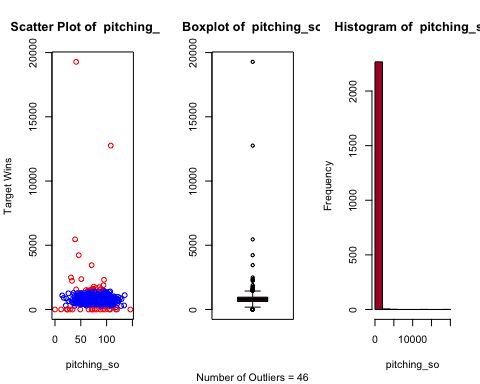
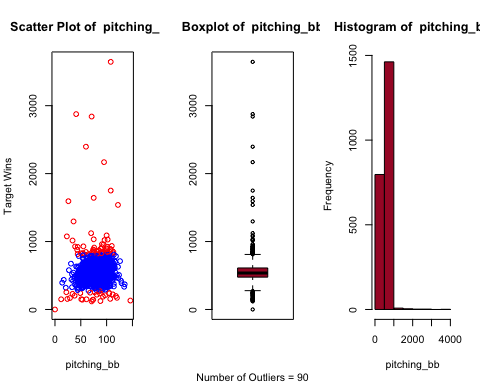
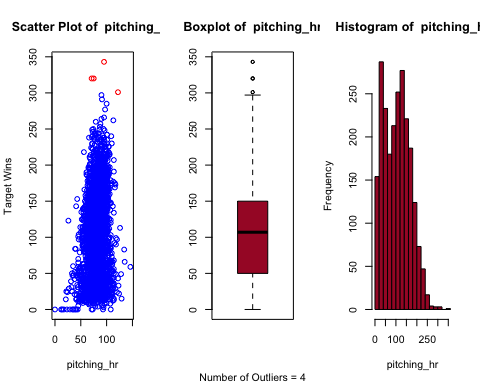
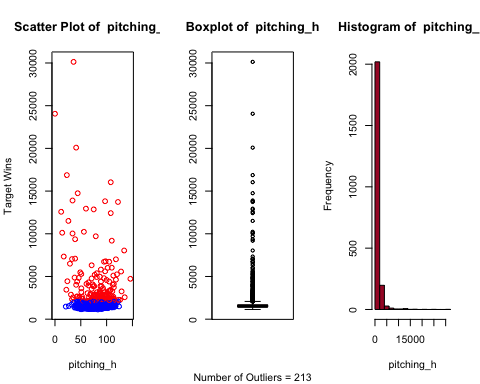
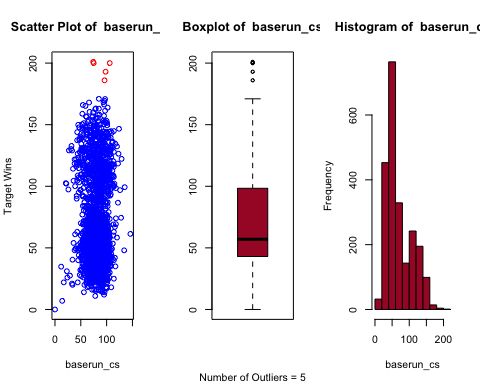
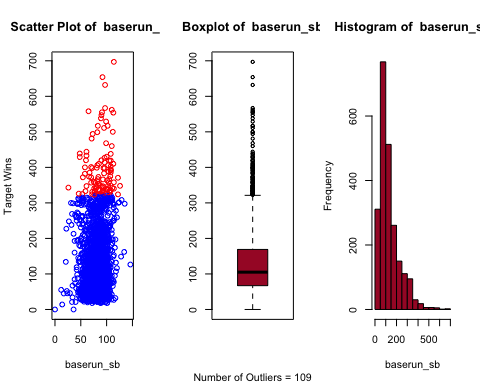
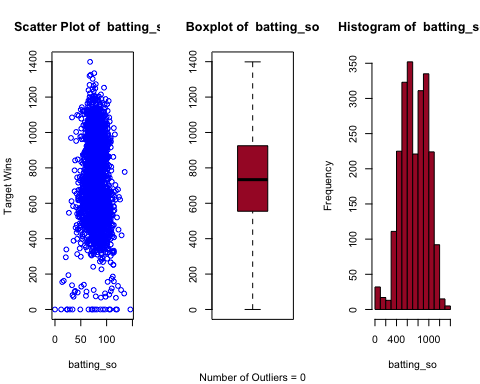
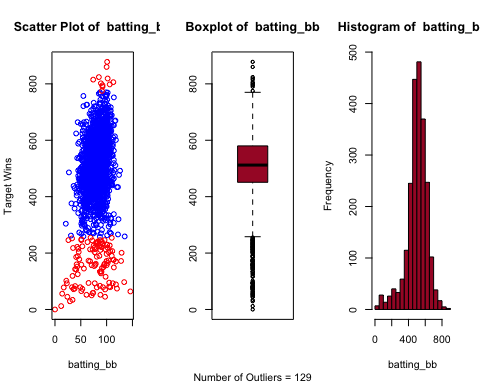
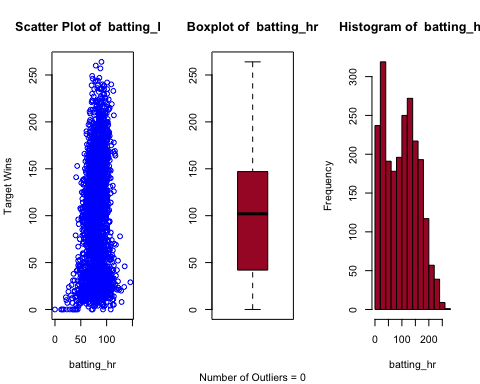
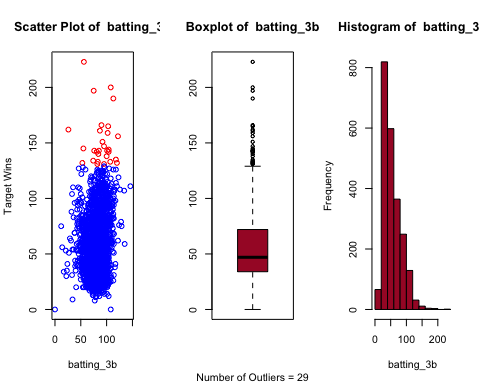
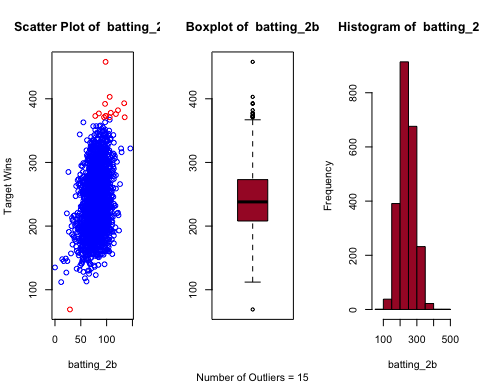
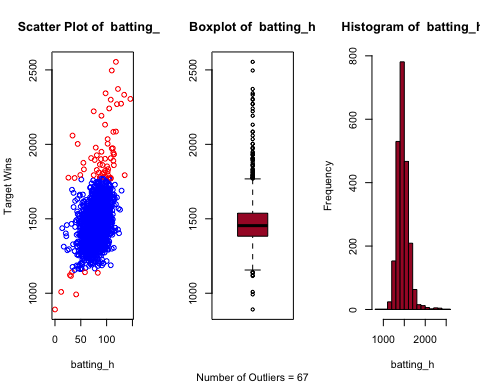
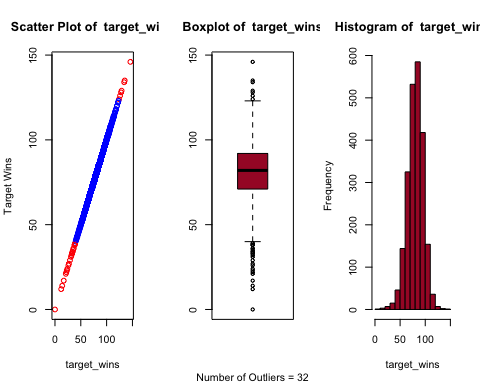
Let's add batting\_hbp back into the data.

moneyball\_MF$batting\_hbp <- if\_else(is.na(batting\_hbp),0,as.numeric(batting\_hbp))  
moneyball\_MF$batting\_hbp\_bi <- batting\_hbp\_bi

### Step 2: Can we find outliers in our Independent and Dependent variables?

Outliers can cause our model to produce the wrong output by influencing its fit. Creating boxplots will aid in identifying those outliers. We can also use the cleveland dotplot to understand the outliers better. This technique uses the row number against actual value to quickly point out any patterns of outliers. This plot will easilly allow us to check the raw data for errors such as typos during the data collection phase. Points on the far right side, or on the far left side, are observed values that are considerably larger, or smaller, than the majority of the observations, and require further investigation. When we use this chart, together with the box plot and histogram, we can easily identify patterns at to where in the data we're seeing outliers.

par(mfrow = c(1, 3))  
i = 2  
while (i %in% c(2:17)) {  
out.lier <- boxplot.stats(moneyball\_MF[,i])$out  
plot(moneyball\_MF$target\_wins, moneyball\_MF[,i],col=ifelse(moneyball\_MF[,i] %in% out.lier, "red", "blue"), xlab = colnames(moneyball\_MF)[i] , ylab = "Target Wins", main = paste("Scatter Plot of ",colnames(moneyball\_MF)[i]))  
   
boxplot(moneyball\_MF[,i], col = "#A71930", main = paste("Boxplot of ",colnames(moneyball\_MF)[i]))  
  
title(sub = paste0("Number of Outliers = ", length(boxplot.stats(moneyball\_MF[,i])$out)))  
  
hist(  
 moneyball\_MF[,i],  
 col = "#A71930",  
 xlab = colnames(moneyball\_MF)[i],  
 main = paste("Histogram of ",colnames(moneyball\_MF)[i])  
)  
 i = i + 1  
}



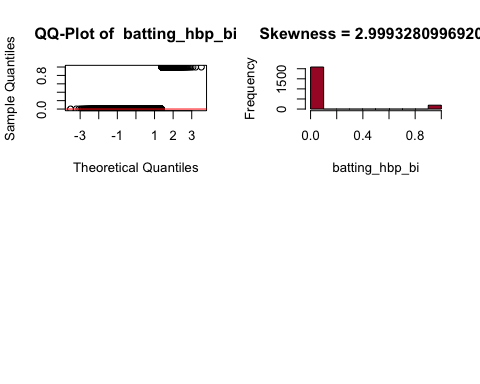
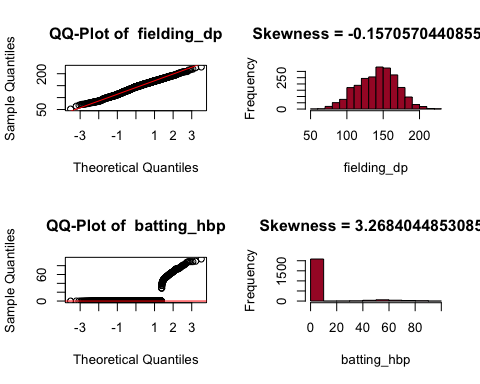
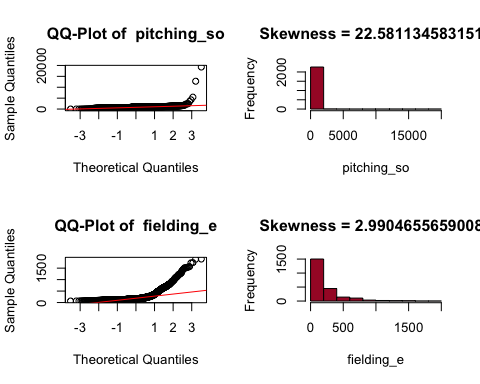
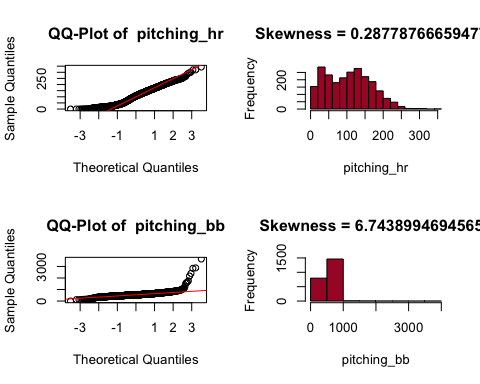
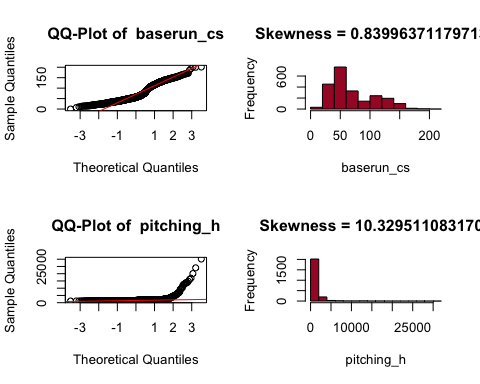
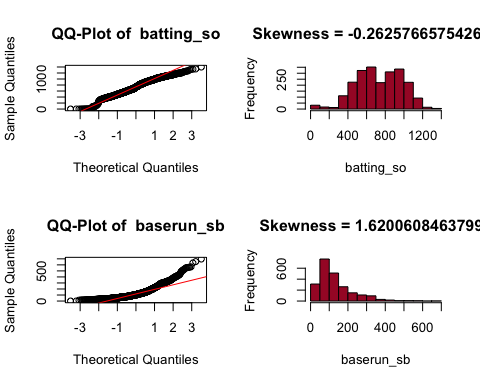
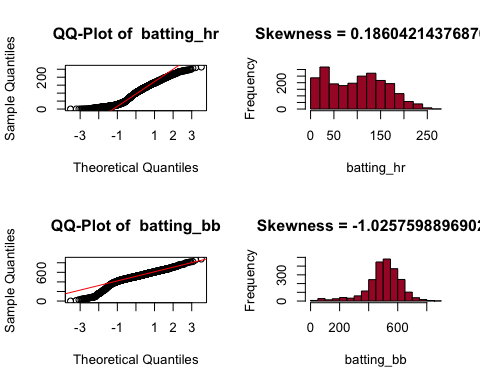
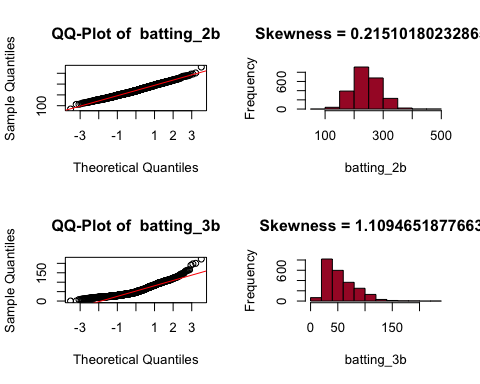
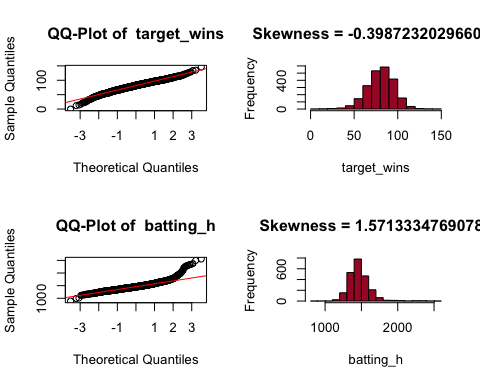
It looks like the outliers are going to be a problem for this model. Multiple techniques will be used to remediate this issue.

Now that step one is done, let's take a look at step 2.

### Step 3: Are the data normally distributed?

From the histogram above, we clearly see the data is not normal, with the exception of some that seems to sort of follow a normal distribution. Let's use QQ-plot to test each column for normality, while adding a histogram and a Skewness number.  
- If skewness is less than −1 or greater than +1, the distribution is highly skewed.  
- If skewness is between −1 and −½ or between +½ and +1, the distribution is moderately skewed.  
- If skewness is between −½ and +½, the distribution is approximately symmetric.

par(mfrow = c(2, 2))  
i = 2  
while (i %in% c(2:18)) {  
 qqnorm(moneyball\_MF[,i], main = paste("QQ-Plot of ",colnames(moneyball\_MF)[i]));qqline(moneyball\_MF[,i], col = 2)  
   
 hist(  
 moneyball\_MF[,i],  
 col = "#A71930",  
 xlab = colnames(moneyball\_MF)[i],  
 main = paste0("Skewness = ",skewness(moneyball\_MF[,i]))  
)  
   
 i = i + 1  
   
}



We would need to try certain transformation to correct for Skewness, with Box-Cox being the number one choice.QQ-plots are a great way to quickly gauge the normality of the variables.

### Step 4: Is there collinearity among the covariates?

Let's create a series of correlation matix to understand how each independent variable interacts with the dependent variable. This correlation matix will help us spot any infrigement of the assupmtions needed to develop a robust OLS model, namely multicollinearity. The caret package can help the user find those pairs and even suggest which one to remove.

The Caret package offers the findcorrelation(), which takes the correlation matrix as an input and finds the fields causing multicollinearity based on a threshold, the cutoff parameter. It in turns returns a vector with values that would need to be removed from our dataset due to correlation.

colnames(moneyball\_MF)[findCorrelation(cor(moneyball\_MF))]

## [1] "batting\_hr" "batting\_hbp"

Per caret's suggestion, we need to remove two variables in order to deal with the multicollinearity issue, batting\_hr and batting\_hbp. We will keep that in mind for when we start the data transformation phase. For now, let's keep them since we need them for more feature engineering. ## Data Transformation

Let's introduce new variables through transformation:

1. batting\_1B = batting\_h-(batting\_2b + batting\_3b + batting\_hr)
2. free\_bases\_num = batting\_hbp + batting\_bb
3. total\_bases = batting\_1B + 2 \* batting\_2b + 3 \* batting\_3b + 4 \* batting\_hr + batting\_bb + batting\_hbp + baserun\_sb
4. total\_bases\_allowed = pitching\_bb + 4 \* pitching\_hr + pitching\_h
5. HR\_over\_OP = batting\_hr - pitching\_hr
6. walks\_over\_OP = batting\_bb - pitching\_bb
7. SO\_over\_OP = pitching\_so - batting\_so

moneyball\_MF$batting\_1B <- moneyball\_MF$batting\_h-(moneyball\_MF$batting\_2b + moneyball\_MF$batting\_3b + moneyball\_MF$batting\_hr)  
moneyball\_MF$free\_bases\_num <- if\_else(is.na(moneyball\_MF$batting\_hbp),0,as.numeric(moneyball\_MF$batting\_hbp)) + moneyball\_MF$batting\_bb  
moneyball\_MF$total\_bases <- moneyball\_MF$batting\_1B + 2 \* moneyball\_MF$batting\_2b + 3 \* moneyball\_MF$batting\_3b + 4 \* moneyball\_MF$batting\_hr + moneyball\_MF$batting\_bb + if\_else(is.na(moneyball\_MF$batting\_hbp),0,as.numeric(moneyball\_MF$batting\_hbp)) + moneyball\_MF$baserun\_sb  
moneyball\_MF$total\_bases\_allowed = moneyball\_MF$pitching\_bb + 4 \* moneyball\_MF$pitching\_hr + moneyball\_MF$pitching\_h  
moneyball\_MF$HR\_over\_OP = moneyball\_MF$batting\_hr - moneyball\_MF$pitching\_hr  
moneyball\_MF$walks\_over\_OP = moneyball\_MF$batting\_bb - moneyball\_MF$pitching\_bb  
moneyball\_MF$SO\_over\_OP = moneyball\_MF$pitching\_so - moneyball\_MF$batting\_so  
# make alist of predictors and format them. This will make it easier when it comes to manually chose variables for the model.  
pred\_list <-  
 "index + target\_wins + batting\_h + batting\_2b + batting\_3b + batting\_hr +  
batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_h + pitching\_hr +  
pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp + batting\_hbp\_bi +  
batting\_1B + free\_bases\_num + total\_bases + total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP + SO\_over\_OP"  
#keep the new variables in a vector for texting later, in cae they don't prove to be of any value.  
new\_var <- c("batting\_1B","free\_bases\_num","total\_bases","total\_bases\_allowed","HR\_over\_OP","walks\_over\_OP","SO\_over\_OP")

Now that we have imputed and created new variables, let's look at the correlation matrix to understand the correlation between the variables and the traget\_wins. Remember when caret suggested to delete batting\_hr and batting\_hbp from our model? Let's build a correlaion matrix to understand why.

moneyball\_MF <- subset(moneyball\_MF, select = -c(batting\_hbp))  
cor(moneyball\_MF)

## index target\_wins batting\_h batting\_2b  
## index 1.000000000 -0.021056435 -0.017920241 0.011183013  
## target\_wins -0.021056435 1.000000000 0.388767521 0.289103645  
## batting\_h -0.017920241 0.388767521 1.000000000 0.562849678  
## batting\_2b 0.011183013 0.289103645 0.562849678 1.000000000  
## batting\_3b -0.005814683 0.142608411 0.427696575 -0.107305824  
## batting\_hr 0.051481047 0.176153200 -0.006544685 0.435397293  
## batting\_bb -0.026567236 0.232559864 -0.072464013 0.255726103  
## batting\_so 0.080827010 -0.035044943 -0.444116757 0.167323725  
## baserun\_sb 0.026442037 0.130226307 0.119420829 -0.194162996  
## baserun\_cs -0.041416296 0.012396284 -0.049331840 -0.388992341  
## pitching\_h 0.017103148 -0.109937054 0.302693709 0.023692188  
## pitching\_hr 0.050985897 0.189013735 0.072853119 0.454550818  
## pitching\_bb -0.015287513 0.124174536 0.094193027 0.178054204  
## pitching\_so 0.054390059 -0.077413116 -0.243587219 0.067125501  
## fielding\_e -0.009233126 -0.176484759 0.264902478 -0.235150986  
## fielding\_dp 0.006045441 0.007788862 0.011908706 0.315467140  
## batting\_hbp\_bi 0.047332196 0.002610647 0.019594018 0.361922796  
## batting\_1B -0.047074417 0.217430135 0.827584756 0.087009889  
## free\_bases\_num -0.019063695 0.228098279 -0.068377971 0.297591911  
## total\_bases 0.022519558 0.481528491 0.627747838 0.705252847  
## total\_bases\_allowed 0.023268954 -0.059959123 0.314205398 0.119290484  
## HR\_over\_OP -0.000553440 -0.060991072 -0.322055891 -0.099453882  
## walks\_over\_OP -0.004745951 0.052184113 -0.162824365 0.011599182  
## SO\_over\_OP 0.019823391 -0.067805571 -0.048131752 -0.008976829  
## batting\_3b batting\_hr batting\_bb batting\_so  
## index -0.005814683 0.051481047 -0.02656724 0.08082701  
## target\_wins 0.142608411 0.176153200 0.23255986 -0.03504494  
## batting\_h 0.427696575 -0.006544685 -0.07246401 -0.44411676  
## batting\_2b -0.107305824 0.435397293 0.25572610 0.16732373  
## batting\_3b 1.000000000 -0.635566946 -0.28723584 -0.66674519  
## batting\_hr -0.635566946 1.000000000 0.51373481 0.71264592  
## batting\_bb -0.287235841 0.513734810 1.00000000 0.38049757  
## batting\_so -0.666745186 0.712645918 0.38049757 1.00000000  
## baserun\_sb 0.553630074 -0.497291546 -0.25992274 -0.26971547  
## baserun\_cs 0.593205169 -0.713034351 -0.35152726 -0.38977962  
## pitching\_h 0.194879411 -0.250145481 -0.44977762 -0.37148536  
## pitching\_hr -0.567836679 0.969371396 0.45955207 0.65360057  
## pitching\_bb -0.002224148 0.136927564 0.48936126 0.04396036  
## pitching\_so -0.260145102 0.185153999 -0.01575343 0.41773397  
## fielding\_e 0.509778447 -0.587339098 -0.65597081 -0.58491326  
## fielding\_dp -0.401839914 0.543730144 0.48949205 0.31934633  
## batting\_hbp\_bi -0.265544426 0.392199209 0.10305838 0.39617487  
## batting\_1B 0.600399234 -0.497294855 -0.35312165 -0.74901850  
## free\_bases\_num -0.316009005 0.553966941 0.99101046 0.42460703  
## total\_bases 0.030030780 0.603209860 0.57208295 0.20430993  
## total\_bases\_allowed 0.092039617 -0.062551344 -0.30004852 -0.24284770  
## HR\_over\_OP -0.243354524 0.074559388 0.19441460 0.20425206  
## walks\_over\_OP -0.231156161 0.266798215 0.27356493 0.26097349  
## SO\_over\_OP 0.043837516 -0.149040175 -0.20564145 -0.03543167  
## baserun\_sb baserun\_cs pitching\_h pitching\_hr  
## index 0.02644204 -0.04141630 0.01710315 0.05098590  
## target\_wins 0.13022631 0.01239628 -0.10993705 0.18901373  
## batting\_h 0.11942083 -0.04933184 0.30269371 0.07285312  
## batting\_2b -0.19416300 -0.38899234 0.02369219 0.45455082  
## batting\_3b 0.55363007 0.59320517 0.19487941 -0.56783668  
## batting\_hr -0.49729155 -0.71303435 -0.25014548 0.96937140  
## batting\_bb -0.25992274 -0.35152726 -0.44977762 0.45955207  
## batting\_so -0.26971547 -0.38977962 -0.37148536 0.65360057  
## baserun\_sb 1.00000000 0.76691081 0.12891566 -0.44994134  
## baserun\_cs 0.76691081 1.00000000 0.05351377 -0.69200215  
## pitching\_h 0.12891566 0.05351377 1.00000000 -0.14161276  
## pitching\_hr -0.44994134 -0.69200215 -0.14161276 1.00000000  
## pitching\_bb 0.07139426 -0.07808740 0.32067616 0.22193750  
## pitching\_so 0.03671354 -0.04680673 0.26740311 0.20493865  
## fielding\_e 0.52259056 0.44425169 0.66775901 -0.49314447  
## fielding\_dp -0.53912041 -0.61100092 -0.25578822 0.51073116  
## batting\_hbp\_bi -0.13606141 -0.24858175 -0.06445004 0.35794984  
## batting\_1B 0.31798518 0.29219507 0.40612014 -0.41549520  
## free\_bases\_num -0.27202750 -0.37632474 -0.44800796 0.49652206  
## total\_bases 0.02626832 -0.31089507 -0.10827230 0.62821797  
## total\_bases\_allowed 0.05671436 -0.07344647 0.97499650 0.05669475  
## HR\_over\_OP -0.16749116 -0.04988651 -0.42822141 -0.17264012  
## walks\_over\_OP -0.29013910 -0.19977181 -0.71949139 0.12897043  
## SO\_over\_OP 0.17386870 0.14142283 0.47798082 -0.09805587  
## pitching\_bb pitching\_so fielding\_e fielding\_dp  
## index -0.015287513 0.054390059 -0.009233126 0.006045441  
## target\_wins 0.124174536 -0.077413116 -0.176484759 0.007788862  
## batting\_h 0.094193027 -0.243587219 0.264902478 0.011908706  
## batting\_2b 0.178054204 0.067125501 -0.235150986 0.315467140  
## batting\_3b -0.002224148 -0.260145102 0.509778447 -0.401839914  
## batting\_hr 0.136927564 0.185153999 -0.587339098 0.543730144  
## batting\_bb 0.489361263 -0.015753433 -0.655970815 0.489492052  
## batting\_so 0.043960357 0.417733972 -0.584913265 0.319346334  
## baserun\_sb 0.071394258 0.036713544 0.522590561 -0.539120409  
## baserun\_cs -0.078087404 -0.046806728 0.444251690 -0.611000919  
## pitching\_h 0.320676162 0.267403113 0.667759010 -0.255788221  
## pitching\_hr 0.221937505 0.204938646 -0.493144466 0.510731164  
## pitching\_bb 1.000000000 0.485086374 -0.022837561 0.148766812  
## pitching\_so 0.485086374 1.000000000 -0.025009441 0.017387359  
## fielding\_e -0.022837561 -0.025009441 1.000000000 -0.562825033  
## fielding\_dp 0.148766812 0.017387359 -0.562825033 1.000000000  
## batting\_hbp\_bi -0.016906833 0.132934373 -0.185315470 0.112376482  
## batting\_1B -0.022820326 -0.328133949 0.547816415 -0.269434497  
## free\_bases\_num 0.476195183 0.002405731 -0.665319984 0.492692295  
## total\_bases 0.361982663 -0.029578841 -0.273572221 0.333320530  
## total\_bases\_allowed 0.459579945 0.346876825 0.557252830 -0.143552266  
## HR\_over\_OP -0.351988418 -0.089553569 -0.353210656 0.106907551  
## walks\_over\_OP -0.704942270 -0.547833641 -0.508313405 0.234023565  
## SO\_over\_OP 0.511809461 0.893197881 0.261972937 -0.138923951  
## batting\_hbp\_bi batting\_1B free\_bases\_num total\_bases  
## index 0.047332196 -0.047074417 -0.019063695 0.02251956  
## target\_wins 0.002610647 0.217430135 0.228098279 0.48152849  
## batting\_h 0.019594018 0.827584756 -0.068377971 0.62774784  
## batting\_2b 0.361922796 0.087009889 0.297591911 0.70525285  
## batting\_3b -0.265544426 0.600399234 -0.316009005 0.03003078  
## batting\_hr 0.392199209 -0.497294855 0.553966941 0.60320986  
## batting\_bb 0.103058382 -0.353121648 0.991010459 0.57208295  
## batting\_so 0.396174874 -0.749018499 0.424607033 0.20430993  
## baserun\_sb -0.136061415 0.317985177 -0.272027496 0.02626832  
## baserun\_cs -0.248581755 0.292195071 -0.376324739 -0.31089507  
## pitching\_h -0.064450039 0.406120142 -0.448007961 -0.10827230  
## pitching\_hr 0.357949841 -0.415495198 0.496522065 0.62821797  
## pitching\_bb -0.016906833 -0.022820326 0.476195183 0.36198266  
## pitching\_so 0.132934373 -0.328133949 0.002405731 -0.02957884  
## fielding\_e -0.185315470 0.547816415 -0.665319984 -0.27357222  
## fielding\_dp 0.112376482 -0.269434497 0.492692295 0.33332053  
## batting\_hbp\_bi 1.000000000 -0.236051718 0.231848863 0.29604889  
## batting\_1B -0.236051718 1.000000000 -0.376395883 0.15822028  
## free\_bases\_num 0.231848863 -0.376395883 1.000000000 0.59846164  
## total\_bases 0.296048893 0.158220276 0.598461643 1.00000000  
## total\_bases\_allowed -0.003909755 0.318513233 -0.293643548 0.04236067  
## HR\_over\_OP 0.119531251 -0.307367362 0.205656303 -0.13158306  
## walks\_over\_OP 0.102464739 -0.262024813 0.280775130 0.06603047  
## SO\_over\_OP -0.049852661 0.009772066 -0.207497853 -0.13365083  
## total\_bases\_allowed HR\_over\_OP walks\_over\_OP  
## index 0.023268954 -0.00055344 -0.004745951  
## target\_wins -0.059959123 -0.06099107 0.052184113  
## batting\_h 0.314205398 -0.32205589 -0.162824365  
## batting\_2b 0.119290484 -0.09945388 0.011599182  
## batting\_3b 0.092039617 -0.24335452 -0.231156161  
## batting\_hr -0.062551344 0.07455939 0.266798215  
## batting\_bb -0.300048525 0.19441460 0.273564933  
## batting\_so -0.242847696 0.20425206 0.260973489  
## baserun\_sb 0.056714362 -0.16749116 -0.290139100  
## baserun\_cs -0.073446473 -0.04988651 -0.199771805  
## pitching\_h 0.974996503 -0.42822141 -0.719491389  
## pitching\_hr 0.056694753 -0.17264012 0.128970430  
## pitching\_bb 0.459579945 -0.35198842 -0.704942270  
## pitching\_so 0.346876825 -0.08955357 -0.547833641  
## fielding\_e 0.557252830 -0.35321066 -0.508313405  
## fielding\_dp -0.143552266 0.10690755 0.234023565  
## batting\_hbp\_bi -0.003909755 0.11953125 0.102464739  
## batting\_1B 0.318513233 -0.30736736 -0.262024813  
## free\_bases\_num -0.293643548 0.20565630 0.280775130  
## total\_bases 0.042360671 -0.13158306 0.066030466  
## total\_bases\_allowed 1.000000000 -0.48106409 -0.750919119  
## HR\_over\_OP -0.481064087 1.00000000 0.546339879  
## walks\_over\_OP -0.750919119 0.54633988 1.000000000  
## SO\_over\_OP 0.501732558 -0.19959081 -0.731743773  
## SO\_over\_OP  
## index 0.019823391  
## target\_wins -0.067805571  
## batting\_h -0.048131752  
## batting\_2b -0.008976829  
## batting\_3b 0.043837516  
## batting\_hr -0.149040175  
## batting\_bb -0.205641453  
## batting\_so -0.035431675  
## baserun\_sb 0.173868698  
## baserun\_cs 0.141422829  
## pitching\_h 0.477980816  
## pitching\_hr -0.098055867  
## pitching\_bb 0.511809461  
## pitching\_so 0.893197881  
## fielding\_e 0.261972937  
## fielding\_dp -0.138923951  
## batting\_hbp\_bi -0.049852661  
## batting\_1B 0.009772066  
## free\_bases\_num -0.207497853  
## total\_bases -0.133650827  
## total\_bases\_allowed 0.501732558  
## HR\_over\_OP -0.199590808  
## walks\_over\_OP -0.731743773  
## SO\_over\_OP 1.000000000

Now that we created new variables, let's see what caret has to say about which variables to remove.

colnames(moneyball\_MF)[findCorrelation(cor(moneyball\_MF), cutoff = 0.9)]

## [1] "batting\_hr" "free\_bases\_num" "pitching\_h"

It suggesting batting\_hr together with free\_bases\_num and pitching\_h. According to the correlation matrix, batting\_hr has a coefficient of correlation of 0.96 related to pitching\_hr, free\_bases\_num has a coefficient of correlation of 0.99 related to batting\_bb, and pitching\_h has a coefficient of correlation of 0.99 related to total\_bases\_allowed. All these variables had a correlation of above th cuttoff point, 0.9. Let's remove those variables.

moneyball\_MF <- subset(moneyball\_MF, select = -c(batting\_hr, free\_bases\_num, pitching\_h))  
  
pred\_list <- "index + target\_wins + batting\_h + batting\_2b + batting\_3b + batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP + SO\_over\_OP"

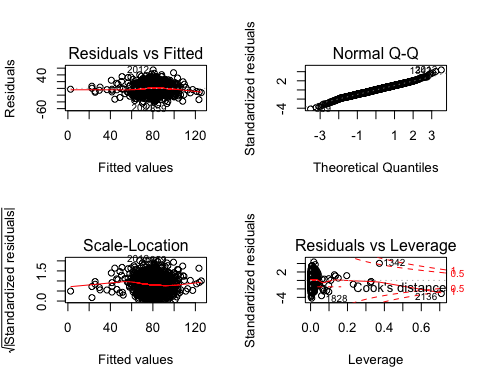
## Build a Model

Let's test a model to establish a baseline

str(moneyball\_MF)

## 'data.frame': 2276 obs. of 21 variables:  
## $ index : num 1 2 3 4 5 6 7 8 11 12 ...  
## $ target\_wins : num 39 70 86 70 82 75 80 85 86 76 ...  
## $ batting\_h : num 1445 1339 1377 1387 1297 ...  
## $ batting\_2b : num 194 219 232 209 186 200 179 171 197 213 ...  
## $ batting\_3b : num 39 22 35 38 27 36 54 37 40 18 ...  
## $ batting\_bb : num 143 685 602 451 472 443 525 456 447 441 ...  
## $ batting\_so : num 842 1075 917 922 920 ...  
## $ baserun\_sb : num 266 37 46 43 49 ...  
## $ baserun\_cs : num 100 28 27 30 39 ...  
## $ pitching\_hr : num 84 191 137 97 102 92 122 116 114 96 ...  
## $ pitching\_bb : num 927 689 602 454 472 443 525 459 447 441 ...  
## $ pitching\_so : num 5456 1082 917 928 920 ...  
## $ fielding\_e : num 1011 193 175 164 138 ...  
## $ fielding\_dp : num 112 155 153 156 168 ...  
## $ batting\_hbp\_bi : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ batting\_1B : num 1199 908 973 1044 982 ...  
## $ total\_bases : num 2165 2894 2738 2454 2364 ...  
## $ total\_bases\_allowed: num 10627 2800 2527 2238 2177 ...  
## $ HR\_over\_OP : num -71 -1 0 -1 0 0 0 -1 0 0 ...  
## $ walks\_over\_OP : num -784 -4 0 -3 0 0 0 -3 0 0 ...  
## $ SO\_over\_OP : num 4614 7 0 6 0 ...

base\_model\_all <- lm(target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP + SO\_over\_OP, data = moneyball\_MF)  
par(mfrow = c(2,2))  
plot(base\_model\_all)



summary(base\_model\_all)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_h + batting\_2b + batting\_3b +   
## batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_hr +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP, data = moneyball\_MF)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -52.652 -8.409 0.127 8.244 54.758   
##   
## Coefficients: (3 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.0416928 5.5994557 5.544 3.31e-08 \*\*\*  
## batting\_h -0.1683958 0.2842424 -0.592 0.5536   
## batting\_2b 0.0551870 0.1442124 0.383 0.7020   
## batting\_3b 0.0128822 0.0765507 0.168 0.8664   
## batting\_bb -0.0622595 0.0709777 -0.877 0.3805   
## batting\_so -0.0147684 0.0025808 -5.723 1.19e-08 \*\*\*  
## baserun\_sb -0.0246874 0.0708188 -0.349 0.7274   
## baserun\_cs 0.0209196 0.0157071 1.332 0.1830   
## pitching\_hr 0.0085637 0.0237397 0.361 0.7183   
## pitching\_bb -0.0039146 0.0042004 -0.932 0.3515   
## pitching\_so 0.0017889 0.0008893 2.012 0.0444 \*   
## fielding\_e -0.0352387 0.0025861 -13.626 < 2e-16 \*\*\*  
## fielding\_dp -0.1256495 0.0139520 -9.006 < 2e-16 \*\*\*  
## batting\_hbp\_bi -8.3004969 4.3301434 -1.917 0.0554 .   
## batting\_1B 0.1384228 0.2139085 0.647 0.5176   
## total\_bases 0.0740487 0.0707177 1.047 0.2952   
## total\_bases\_allowed 0.0004301 0.0003704 1.161 0.2457   
## HR\_over\_OP NA NA NA NA   
## walks\_over\_OP NA NA NA NA   
## SO\_over\_OP NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.62 on 2259 degrees of freedom  
## Multiple R-squared: 0.3623, Adjusted R-squared: 0.3578   
## F-statistic: 80.21 on 16 and 2259 DF, p-value: < 2.2e-16

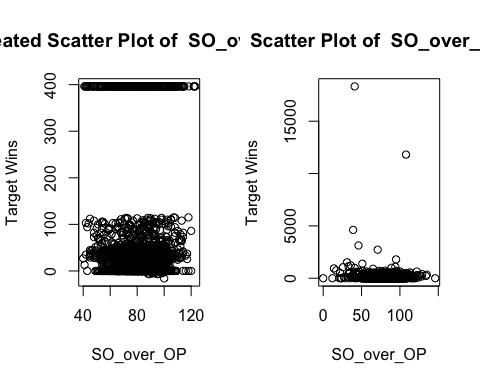
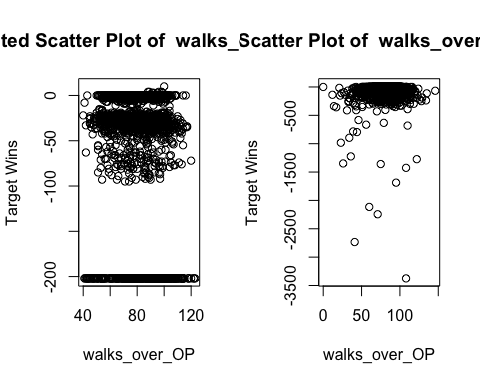
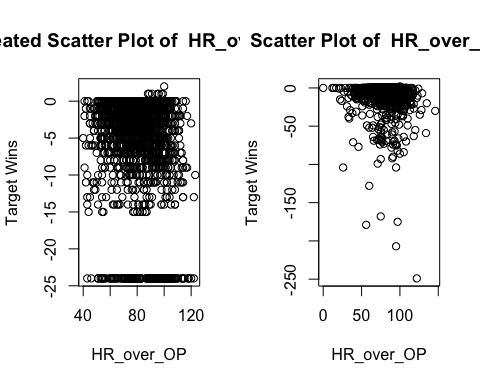
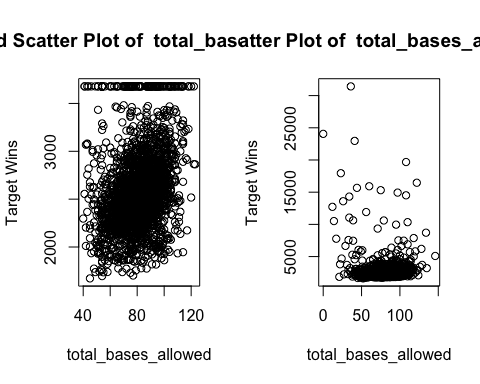
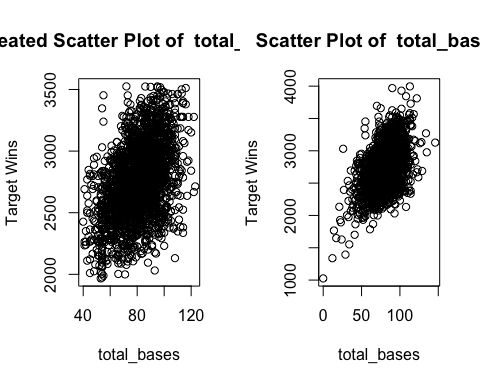
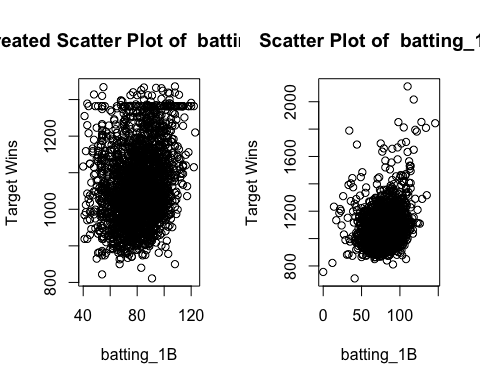
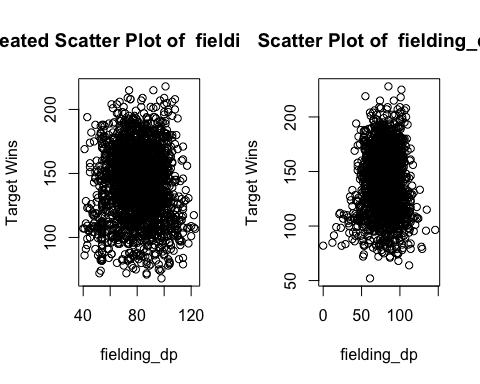
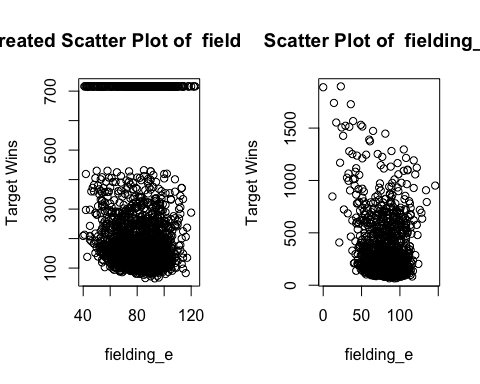
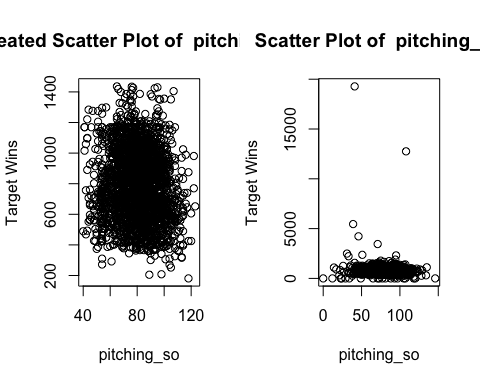
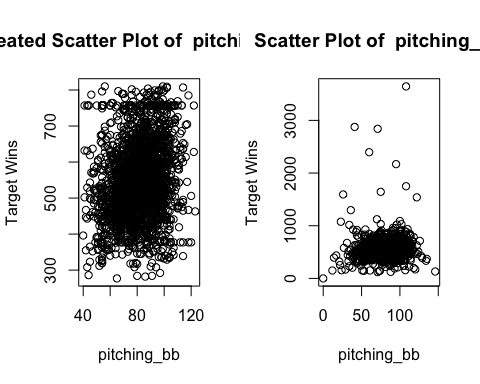
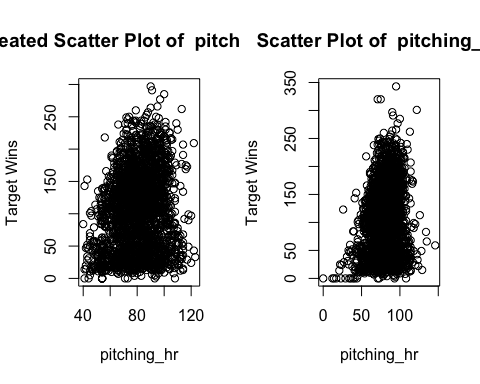
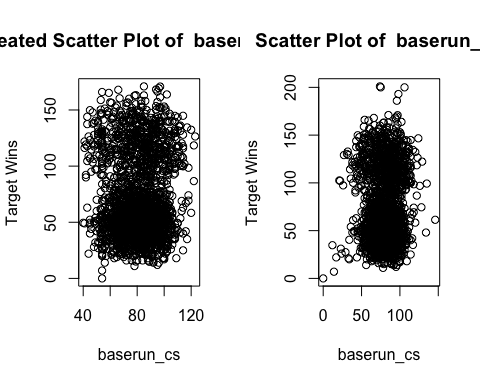
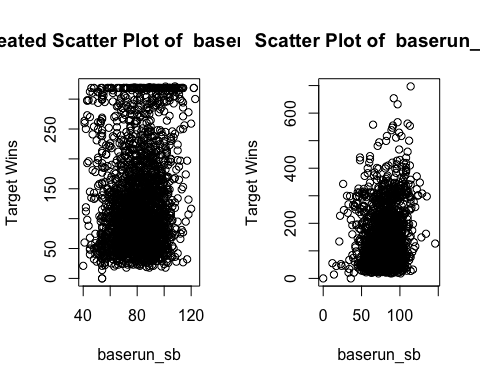
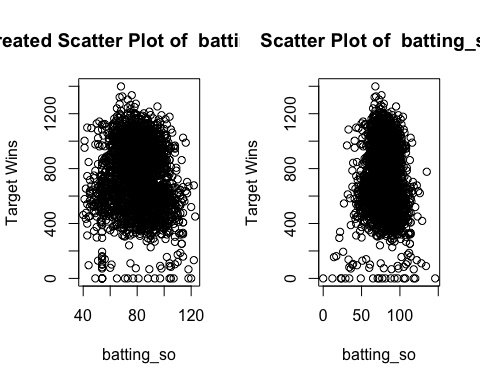
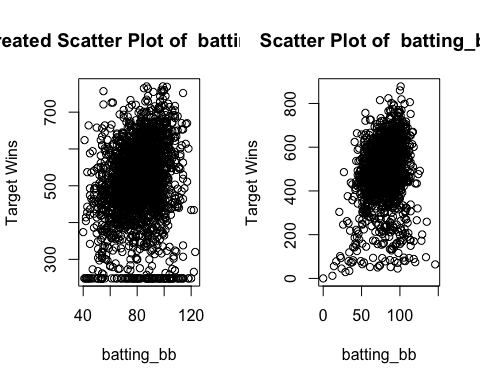
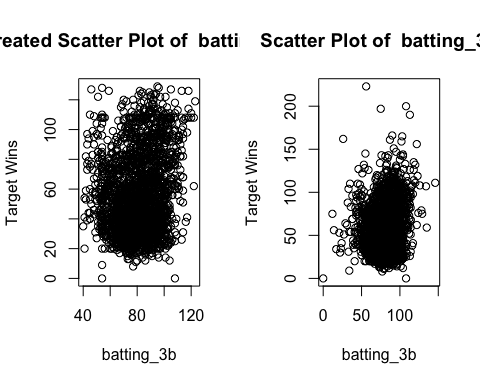
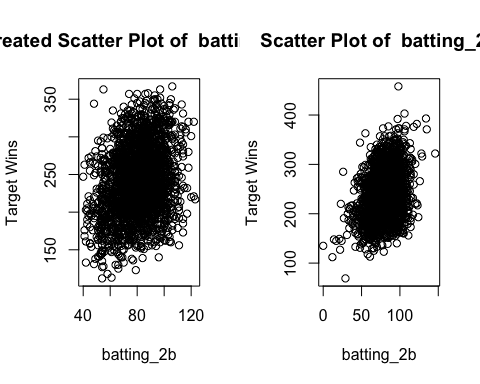
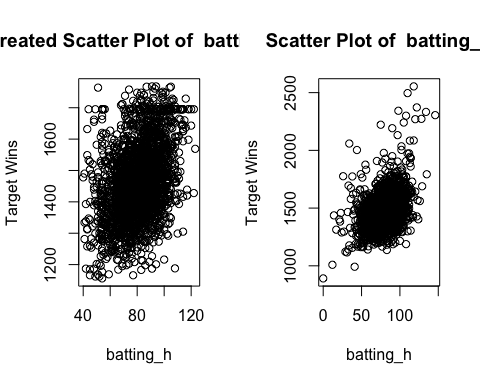
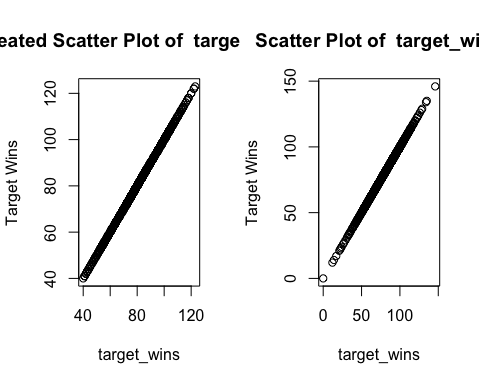
mse <- function(sm)   
 mean(sm$residuals^2)  
  
paste('MSE equal ', mse(base\_model\_all), "and RMSE is ", sqrt(mse(base\_model\_all)))

## [1] "MSE equal 158.166007039155 and RMSE is 12.576406761836"

Though R-squared and adjusted R-square is decent, we can clearly see that this model is not optmal. Let's try to forget about the new additions, and build a model without them.

Let's fix the issue with outliers and see if we get any improvements. For the first approach we will use Winsoring approch.  
For every outlier we will impute it with Q1 - 1.5\*IQR or Q3 + 1.5\*IQR, the cutoff for outliers.

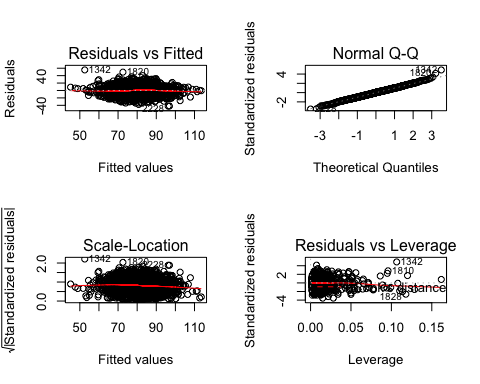
outlier\_treat <- moneyball\_MF[,-c(1,15)]  
comp\_data <- moneyball\_MF[,-c(1,15)]  
i = 1  
while (i %in% seq\_along(outlier\_treat)) {  
  
qnt <- quantile(outlier\_treat[,i], probs = c(.25, .75), na.rm = T)  
caps <- quantile(outlier\_treat[,i], probs = c(.05, .95), na.rm = T)  
H <- 1.5 \* IQR(outlier\_treat[,i], na.rm = T)  
outlier\_treat[,i][outlier\_treat[,i] < (qnt[1] - H)] <- caps[1]  
outlier\_treat[,i][outlier\_treat[,i] > (qnt[2] + H)] <- caps[2]  
 par(mfrow = c(1,2))   
 plot(outlier\_treat$target\_wins, outlier\_treat[,i], xlab = colnames(outlier\_treat)[i] , ylab = "Target Wins", main = paste("Treated Scatter Plot of ",colnames(outlier\_treat)[i]))  
 plot(comp\_data$target\_wins, comp\_data[,i],xlab = colnames(comp\_data)[i] , ylab = "Target Wins", main = paste("Scatter Plot of ",colnames(comp\_data)[i]))  
 i = i + 1  
}



#add back the columns that we dropped prior to the outlier treatment  
outlier\_treat <- cbind(outlier\_treat,moneyball\_MF[,c(1,15)])

Let's try different models using the new data.

base\_model\_orig <-  
 lm(target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP + SO\_over\_OP, data = outlier\_treat)  
 par(mfrow = c(2, 2))  
 plot(base\_model\_orig)



summary(base\_model\_orig)

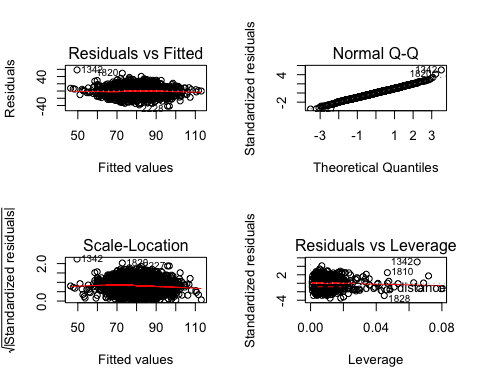
##   
## Call:  
## lm(formula = target\_wins ~ batting\_h + batting\_2b + batting\_3b +   
## batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_hr +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP, data = outlier\_treat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -42.233 -8.088 0.253 7.745 55.427   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 35.123619 6.138975 5.721 1.20e-08 \*\*\*  
## batting\_h 0.004055 0.013627 0.298 0.76608   
## batting\_2b -0.035640 0.014628 -2.437 0.01491 \*   
## batting\_3b 0.054027 0.022752 2.375 0.01765 \*   
## batting\_bb 0.026962 0.009036 2.984 0.00288 \*\*   
## batting\_so -0.011793 0.005336 -2.210 0.02721 \*   
## baserun\_sb 0.063622 0.008905 7.145 1.21e-12 \*\*\*  
## baserun\_cs -0.001510 0.017519 -0.086 0.93134   
## pitching\_hr -0.013255 0.018338 -0.723 0.46987   
## pitching\_bb -0.035640 0.007652 -4.658 3.38e-06 \*\*\*  
## pitching\_so -0.005737 0.004952 -1.159 0.24674   
## fielding\_e -0.042457 0.003071 -13.827 < 2e-16 \*\*\*  
## fielding\_dp -0.098734 0.013254 -7.449 1.33e-13 \*\*\*  
## batting\_hbp\_bi -4.420561 1.132281 -3.904 9.73e-05 \*\*\*  
## batting\_1B -0.009107 0.012942 -0.704 0.48173   
## total\_bases 0.022028 0.004502 4.893 1.06e-06 \*\*\*  
## total\_bases\_allowed 0.012397 0.002050 6.047 1.72e-09 \*\*\*  
## HR\_over\_OP 0.010129 0.081686 0.124 0.90132   
## walks\_over\_OP 0.034089 0.010948 3.114 0.00187 \*\*   
## SO\_over\_OP 0.012261 0.004447 2.757 0.00588 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12 on 2256 degrees of freedom  
## Multiple R-squared: 0.3459, Adjusted R-squared: 0.3404   
## F-statistic: 62.8 on 19 and 2256 DF, p-value: < 2.2e-16

paste('MSE equal ', mse(base\_model\_orig))

## [1] "MSE equal 142.763337742595"

This model looks good, from a performance point of view(r2), but when I look at the variance of the residual I don't feel secure. Specially after analyising Cook's distance graph. There are several observations that are way out from the rest. Let's build another model including only those with low p-Values.

base\_model\_lp <-  
 lm(target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so + baserun\_sb + pitching\_bb +   
 fielding\_e + fielding\_dp + batting\_hbp\_bi + total\_bases + total\_bases\_allowed + walks\_over\_OP + SO\_over\_OP, data = outlier\_treat)  
 par(mfrow = c(2, 2))  
 plot(base\_model\_lp)



summary(base\_model\_lp)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + pitching\_bb + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + total\_bases + total\_bases\_allowed + walks\_over\_OP +   
## SO\_over\_OP, data = outlier\_treat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -41.832 -8.064 0.175 7.833 58.399   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 32.042515 3.269463 9.801 < 2e-16 \*\*\*  
## batting\_2b -0.028986 0.009041 -3.206 0.001364 \*\*   
## batting\_3b 0.058833 0.017710 3.322 0.000908 \*\*\*  
## batting\_bb 0.030852 0.008469 3.643 0.000276 \*\*\*  
## batting\_so -0.016710 0.001804 -9.261 < 2e-16 \*\*\*  
## baserun\_sb 0.063457 0.005851 10.845 < 2e-16 \*\*\*  
## pitching\_bb -0.036726 0.007442 -4.935 8.59e-07 \*\*\*  
## fielding\_e -0.041706 0.002934 -14.214 < 2e-16 \*\*\*  
## fielding\_dp -0.099706 0.012883 -7.739 1.50e-14 \*\*\*  
## batting\_hbp\_bi -4.226860 1.092614 -3.869 0.000113 \*\*\*  
## total\_bases 0.020971 0.002438 8.602 < 2e-16 \*\*\*  
## total\_bases\_allowed 0.011086 0.001513 7.326 3.29e-13 \*\*\*  
## walks\_over\_OP 0.034041 0.010551 3.226 0.001272 \*\*   
## SO\_over\_OP 0.010053 0.003898 2.579 0.009966 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.99 on 2262 degrees of freedom  
## Multiple R-squared: 0.3452, Adjusted R-squared: 0.3414   
## F-statistic: 91.72 on 13 and 2262 DF, p-value: < 2.2e-16

paste('MSE equal ', mse(base\_model\_lp))

## [1] "MSE equal 142.927674156311"

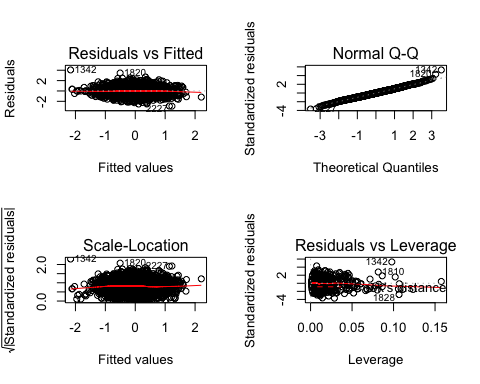
Though the rsquared value went down, there are some improvements on the Cook's distance chart. Now let's try to use use the caret package to apply the transformations we discussed earlier in our exploration phase. I will include all the variables minus the ones cause Multicollinearity issues.

1. Center and Scale the data
2. Fix the the problem with outliers by using spatial sign Transformation
3. Last but not least a boxcox transformation to take car of the skewness

trans <- preProcess(outlier\_treat, method = c("BoxCox","center", "scale"))  
transformed <- predict(trans, outlier\_treat)  
head(transformed)

## target\_wins batting\_h batting\_2b batting\_3b batting\_bb batting\_so  
## 1 -1.76727657 -0.1111056 -1.0260875 -0.5976418 -2.0900610 0.4509070  
## 2 -0.76118502 -1.0523406 -0.4551888 -1.2453898 1.8582866 1.4070416  
## 3 0.31687312 -0.7066133 -0.1661113 -0.7500531 0.9112549 0.7586757  
## 4 -0.76118502 -0.6172179 -0.6810714 -0.6357446 -0.5850077 0.7791936  
## 5 0.04120325 -1.4460647 -1.2133573 -1.0548757 -0.3951198 0.7709864  
## 6 -0.43149374 -1.6187221 -0.8871532 -0.7119502 -0.6557485 0.9884763  
## baserun\_sb baserun\_cs pitching\_hr pitching\_bb pitching\_so fielding\_e  
## 1 1.6766913 0.8365039 -0.35375893 1.8815144 1.6531432 1.73665912  
## 2 -1.1245933 -1.1553448 1.40662888 1.3308308 1.2673652 0.25915028  
## 3 -1.0147342 -1.1829825 0.51820887 0.5763990 0.5378137 0.06267378  
## 4 -1.0513539 -1.1000694 -0.13988003 -0.8749743 0.5864504 -0.07748718  
## 5 -0.9781145 -0.8513302 -0.05761892 -0.6845719 0.5510782 -0.49255669  
## 6 -0.2701337 -0.2985763 -0.22214115 -0.9935698 0.7854190 -0.80732240  
## fielding\_dp batting\_1B total\_bases total\_bases\_allowed HR\_over\_OP  
## 1 -1.1301341 1.2688903 -2.02196615 2.11040379 -2.9180993  
## 2 0.4543355 -1.7915976 0.45456779 0.59390081 0.5315340  
## 3 0.3777646 -0.9158042 -0.07574655 -0.05571369 0.6815181  
## 4 0.4927324 -0.1066855 -1.04119059 -0.88758315 0.5315340  
## 5 0.9591689 -0.8054425 -1.34714116 -1.08694911 0.6815181  
## 6 0.2255249 -1.1961690 -1.30294830 -1.38831137 0.6815181  
## walks\_over\_OP SO\_over\_OP index batting\_hbp\_bi  
## 1 -2.2942806 2.5117883 -2.185553 -0.3025995  
## 2 0.6102467 -0.5049697 -2.175875 -0.3025995  
## 3 0.6689240 -0.5592210 -2.167613 -0.3025995  
## 4 0.6249160 -0.5127199 -2.160153 -0.3025995  
## 5 0.6689240 -0.5592210 -2.153239 -0.3025995  
## 6 0.6689240 -0.5592210 -2.146731 -0.3025995

trans\_model\_all <-  
 lm(target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP + SO\_over\_OP, data = transformed)  
 par(mfrow = c(2, 2))  
 plot(trans\_model\_all)



summary(trans\_model\_all)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_h + batting\_2b + batting\_3b +   
## batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_hr +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP, data = transformed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.9324 -0.5430 -0.0133 0.5176 4.0620   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.343e-13 1.694e-02 0.000 1.000000   
## batting\_h -4.445e-02 1.009e-01 -0.441 0.659536   
## batting\_2b -4.135e-02 4.174e-02 -0.991 0.321924   
## batting\_3b 1.657e-01 3.972e-02 4.171 3.14e-05 \*\*\*  
## batting\_bb 1.950e-01 5.587e-02 3.490 0.000492 \*\*\*  
## batting\_so -2.226e-01 8.274e-02 -2.690 0.007200 \*\*   
## baserun\_sb 1.882e-01 4.701e-02 4.002 6.48e-05 \*\*\*  
## baserun\_cs 1.472e-01 4.353e-02 3.382 0.000733 \*\*\*  
## pitching\_hr 1.491e-02 7.246e-02 0.206 0.836935   
## pitching\_bb -1.950e-01 4.696e-02 -4.152 3.42e-05 \*\*\*  
## pitching\_so -9.011e-02 7.167e-02 -1.257 0.208772   
## fielding\_e -5.770e-01 3.884e-02 -14.858 < 2e-16 \*\*\*  
## fielding\_dp -1.951e-01 2.375e-02 -8.216 3.52e-16 \*\*\*  
## batting\_hbp\_bi -1.375e-01 2.145e-02 -6.413 1.73e-10 \*\*\*  
## batting\_1B 4.819e-02 7.787e-02 0.619 0.536041   
## total\_bases 3.530e-01 8.664e-02 4.075 4.77e-05 \*\*\*  
## total\_bases\_allowed 2.413e-01 5.831e-02 4.138 3.64e-05 \*\*\*  
## HR\_over\_OP -5.382e-02 3.604e-02 -1.493 0.135467   
## walks\_over\_OP 2.077e-01 5.003e-02 4.151 3.43e-05 \*\*\*  
## SO\_over\_OP 6.148e-02 3.833e-02 1.604 0.108868   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.808 on 2256 degrees of freedom  
## Multiple R-squared: 0.3526, Adjusted R-squared: 0.3472   
## F-statistic: 64.68 on 19 and 2256 DF, p-value: < 2.2e-16

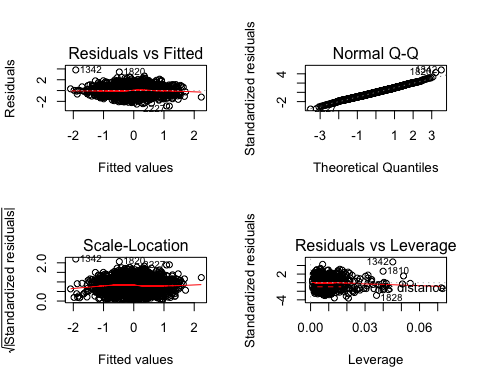
paste('MSE equal ', mse(trans\_model\_all))

## [1] "MSE equal 0.647083867565306"

The residual plots look pretty good, with the exception of some possibly influential observation. Looking at Cook's Distance, it's clear that we have influential data, but the other charts look right where they should be.

Let's look at another model using the same transformed data, but now looking only on the columns with low p-value.

trans\_model\_lp <-  
 lm(target\_wins ~ batting\_3b + batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e + fielding\_dp + batting\_hbp\_bi + total\_bases + total\_bases\_allowed + walks\_over\_OP + SO\_over\_OP, data = transformed)  
 par(mfrow = c(2, 2))  
 plot(trans\_model\_lp)



summary(trans\_model\_lp)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + total\_bases + total\_bases\_allowed + walks\_over\_OP +   
## SO\_over\_OP, data = transformed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8851 -0.5568 -0.0115 0.5256 3.8180   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.213e-15 1.695e-02 0.000 1.00000   
## batting\_3b 1.711e-01 3.287e-02 5.203 2.13e-07 \*\*\*  
## batting\_bb 2.165e-01 5.059e-02 4.280 1.94e-05 \*\*\*  
## batting\_so -3.174e-01 3.045e-02 -10.426 < 2e-16 \*\*\*  
## baserun\_sb 1.919e-01 4.007e-02 4.790 1.77e-06 \*\*\*  
## baserun\_cs 1.388e-01 4.267e-02 3.253 0.00116 \*\*   
## pitching\_bb -1.972e-01 4.608e-02 -4.278 1.96e-05 \*\*\*  
## fielding\_e -5.694e-01 3.797e-02 -14.996 < 2e-16 \*\*\*  
## fielding\_dp -1.903e-01 2.341e-02 -8.129 7.06e-16 \*\*\*  
## batting\_hbp\_bi -1.529e-01 2.027e-02 -7.545 6.53e-14 \*\*\*  
## total\_bases 3.296e-01 4.125e-02 7.990 2.12e-15 \*\*\*  
## total\_bases\_allowed 2.218e-01 4.262e-02 5.204 2.12e-07 \*\*\*  
## walks\_over\_OP 1.903e-01 4.811e-02 3.955 7.89e-05 \*\*\*  
## SO\_over\_OP 6.804e-02 3.372e-02 2.018 0.04374 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8086 on 2262 degrees of freedom  
## Multiple R-squared: 0.3499, Adjusted R-squared: 0.3462   
## F-statistic: 93.65 on 13 and 2262 DF, p-value: < 2.2e-16

paste('MSE equal ', mse(trans\_model\_lp))

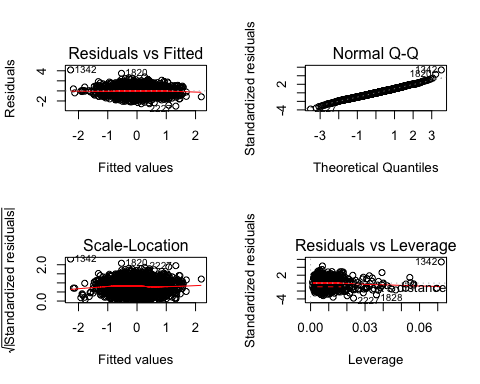
## [1] "MSE equal 0.649826477602929"

This model seems to be on par with the other models. I'll try stepwise approaches and then I'll see if removing "influencial" observations will improve the model. Let's try, stepwise approach.  
1. Both direction

stepwise\_base\_model\_bd <- stepAIC(trans\_model\_all, direction = "both")

## Start: AIC=-950.7  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb +   
## pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - pitching\_hr 1 0.028 1472.8 -952.65  
## - batting\_h 1 0.127 1472.9 -952.50  
## - batting\_1B 1 0.250 1473.0 -952.31  
## - batting\_2b 1 0.641 1473.4 -951.71  
## - pitching\_so 1 1.032 1473.8 -951.10  
## <none> 1472.8 -950.70  
## - HR\_over\_OP 1 1.456 1474.2 -950.45  
## - SO\_over\_OP 1 1.679 1474.4 -950.10  
## - batting\_so 1 4.724 1477.5 -945.41  
## - baserun\_cs 1 7.465 1480.2 -941.19  
## - batting\_bb 1 7.952 1480.7 -940.44  
## - baserun\_sb 1 10.456 1483.2 -936.59  
## - total\_bases 1 10.838 1483.6 -936.01  
## - total\_bases\_allowed 1 11.176 1483.9 -935.49  
## - walks\_over\_OP 1 11.250 1484.0 -935.38  
## - pitching\_bb 1 11.255 1484.0 -935.37  
## - batting\_3b 1 11.358 1484.1 -935.21  
## - batting\_hbp\_bi 1 26.848 1499.6 -911.58  
## - fielding\_dp 1 44.062 1516.8 -885.60  
## - fielding\_e 1 144.108 1616.9 -740.23  
##   
## Step: AIC=-952.65  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_bb + pitching\_so +   
## fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B +   
## total\_bases + total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP +   
## SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - batting\_h 1 0.114 1472.9 -954.48  
## - batting\_1B 1 0.223 1473.0 -954.31  
## - pitching\_so 1 1.016 1473.8 -953.08  
## - batting\_2b 1 1.073 1473.9 -953.00  
## <none> 1472.8 -952.65  
## - SO\_over\_OP 1 1.684 1474.5 -952.05  
## - HR\_over\_OP 1 1.935 1474.7 -951.66  
## + pitching\_hr 1 0.028 1472.8 -950.70  
## - batting\_so 1 4.696 1477.5 -947.41  
## - baserun\_cs 1 7.640 1480.4 -942.88  
## - batting\_bb 1 8.430 1481.2 -941.66  
## - pitching\_bb 1 11.290 1484.1 -937.27  
## - total\_bases\_allowed 1 11.433 1484.2 -937.05  
## - walks\_over\_OP 1 11.596 1484.4 -936.80  
## - baserun\_sb 1 12.371 1485.2 -935.61  
## - batting\_3b 1 13.795 1486.6 -933.43  
## - total\_bases 1 20.707 1493.5 -922.88  
## - batting\_hbp\_bi 1 28.450 1501.2 -911.11  
## - fielding\_dp 1 44.235 1517.0 -887.30  
## - fielding\_e 1 149.796 1622.6 -734.19  
##   
## Step: AIC=-954.48  
## target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + baserun\_cs + pitching\_bb + pitching\_so + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + batting\_1B + total\_bases +   
## total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP + SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - batting\_1B 1 0.151 1473.1 -956.24  
## - pitching\_so 1 0.945 1473.8 -955.02  
## <none> 1472.9 -954.48  
## - SO\_over\_OP 1 1.713 1474.6 -953.83  
## - HR\_over\_OP 1 1.999 1474.9 -953.39  
## + batting\_h 1 0.114 1472.8 -952.65  
## + pitching\_hr 1 0.015 1472.9 -952.50  
## - batting\_2b 1 2.643 1475.5 -952.40  
## - batting\_so 1 5.038 1477.9 -948.70  
## - baserun\_cs 1 7.574 1480.5 -944.80  
## - batting\_bb 1 9.108 1482.0 -942.44  
## - pitching\_bb 1 11.207 1484.1 -939.22  
## - walks\_over\_OP 1 11.516 1484.4 -938.75  
## - total\_bases\_allowed 1 13.437 1486.3 -935.81  
## - baserun\_sb 1 13.967 1486.9 -935.00  
## - batting\_3b 1 14.458 1487.4 -934.24  
## - total\_bases 1 26.309 1499.2 -916.18  
## - batting\_hbp\_bi 1 28.527 1501.4 -912.82  
## - fielding\_dp 1 44.141 1517.0 -889.27  
## - fielding\_e 1 149.682 1622.6 -736.19  
##   
## Step: AIC=-956.24  
## target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + baserun\_cs + pitching\_bb + pitching\_so + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP + walks\_over\_OP + SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - pitching\_so 1 0.982 1474.0 -956.73  
## <none> 1473.1 -956.24  
## - SO\_over\_OP 1 1.794 1474.8 -955.47  
## - HR\_over\_OP 1 1.933 1475.0 -955.26  
## + batting\_1B 1 0.151 1472.9 -954.48  
## + batting\_h 1 0.042 1473.0 -954.31  
## + pitching\_hr 1 0.011 1473.0 -954.26  
## - batting\_2b 1 2.677 1475.7 -954.11  
## - batting\_so 1 5.737 1478.8 -949.40  
## - baserun\_cs 1 7.950 1481.0 -945.99  
## - batting\_bb 1 8.967 1482.0 -944.43  
## - pitching\_bb 1 11.376 1484.4 -940.73  
## - walks\_over\_OP 1 11.520 1484.6 -940.51  
## - total\_bases\_allowed 1 13.510 1486.6 -937.46  
## - baserun\_sb 1 13.917 1487.0 -936.84  
## - batting\_3b 1 14.474 1487.5 -935.99  
## - total\_bases 1 29.324 1502.4 -913.38  
## - batting\_hbp\_bi 1 29.810 1502.9 -912.64  
## - fielding\_dp 1 44.052 1517.1 -891.18  
## - fielding\_e 1 149.533 1622.6 -738.19  
##   
## Step: AIC=-956.73  
## target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - SO\_over\_OP 1 1.218 1475.3 -956.85  
## <none> 1474.0 -956.73  
## + pitching\_so 1 0.982 1473.1 -956.24  
## - HR\_over\_OP 1 1.866 1475.9 -955.85  
## + batting\_1B 1 0.188 1473.8 -955.02  
## + batting\_h 1 0.091 1474.0 -954.87  
## + pitching\_hr 1 0.028 1474.0 -954.77  
## - batting\_2b 1 2.989 1477.0 -954.12  
## - baserun\_cs 1 7.544 1481.6 -947.11  
## - batting\_bb 1 9.823 1483.9 -943.61  
## - walks\_over\_OP 1 11.345 1485.4 -941.28  
## - pitching\_bb 1 12.096 1486.1 -940.13  
## - baserun\_sb 1 13.390 1487.4 -938.14  
## - total\_bases\_allowed 1 13.685 1487.7 -937.69  
## - batting\_3b 1 14.198 1488.2 -936.91  
## - batting\_hbp\_bi 1 30.730 1504.8 -911.77  
## - total\_bases 1 40.213 1514.2 -897.47  
## - fielding\_dp 1 43.282 1517.3 -892.86  
## - batting\_so 1 73.325 1547.4 -848.23  
## - fielding\_e 1 148.828 1622.9 -739.80  
##   
## Step: AIC=-956.85  
## target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## <none> 1475.3 -956.85  
## + SO\_over\_OP 1 1.218 1474.0 -956.73  
## + pitching\_so 1 0.406 1474.8 -955.47  
## + batting\_1B 1 0.248 1475.0 -955.23  
## + batting\_h 1 0.107 1475.2 -955.01  
## + pitching\_hr 1 0.034 1475.2 -954.90  
## - batting\_2b 1 2.697 1478.0 -954.69  
## - HR\_over\_OP 1 3.776 1479.0 -953.03  
## - baserun\_cs 1 8.833 1484.1 -945.26  
## - batting\_bb 1 9.329 1484.6 -944.50  
## - walks\_over\_OP 1 10.136 1485.4 -943.26  
## - pitching\_bb 1 11.722 1487.0 -940.83  
## - baserun\_sb 1 13.569 1488.8 -938.01  
## - batting\_3b 1 14.061 1489.3 -937.26  
## - total\_bases\_allowed 1 14.737 1490.0 -936.22  
## - batting\_hbp\_bi 1 30.884 1506.1 -911.69  
## - total\_bases 1 38.995 1514.2 -899.47  
## - fielding\_dp 1 43.296 1518.5 -893.01  
## - batting\_so 1 74.474 1549.7 -846.75  
## - fielding\_e 1 148.669 1623.9 -740.32

par(mfrow = c(2, 2))  
 plot(stepwise\_base\_model\_bd)



summary(stepwise\_base\_model\_bd)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP + walks\_over\_OP, data = transformed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.0325 -0.5448 -0.0108 0.5230 4.1704   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.764e-15 1.693e-02 0.000 1.000000   
## batting\_2b -5.547e-02 2.728e-02 -2.033 0.042159 \*   
## batting\_3b 1.549e-01 3.337e-02 4.642 3.64e-06 \*\*\*  
## batting\_bb 1.950e-01 5.158e-02 3.781 0.000160 \*\*\*  
## batting\_so -3.124e-01 2.924e-02 -10.684 < 2e-16 \*\*\*  
## baserun\_sb 1.851e-01 4.059e-02 4.560 5.38e-06 \*\*\*  
## baserun\_cs 1.552e-01 4.219e-02 3.679 0.000239 \*\*\*  
## pitching\_bb -1.952e-01 4.605e-02 -4.239 2.34e-05 \*\*\*  
## fielding\_e -5.754e-01 3.812e-02 -15.095 < 2e-16 \*\*\*  
## fielding\_dp -1.910e-01 2.345e-02 -8.146 6.15e-16 \*\*\*  
## batting\_hbp\_bi -1.423e-01 2.068e-02 -6.880 7.72e-12 \*\*\*  
## total\_bases 3.721e-01 4.813e-02 7.731 1.60e-14 \*\*\*  
## total\_bases\_allowed 2.054e-01 4.323e-02 4.752 2.14e-06 \*\*\*  
## HR\_over\_OP -7.286e-02 3.029e-02 -2.406 0.016226 \*   
## walks\_over\_OP 1.805e-01 4.580e-02 3.941 8.35e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8078 on 2261 degrees of freedom  
## Multiple R-squared: 0.3515, Adjusted R-squared: 0.3475   
## F-statistic: 87.55 on 14 and 2261 DF, p-value: < 2.2e-16

paste('MSE equal ', mse(stepwise\_base\_model\_bd))

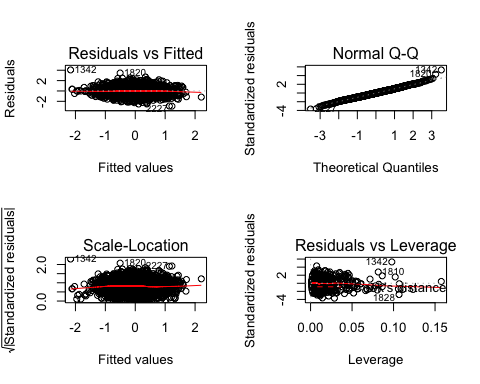
## [1] "MSE equal 0.648179170553685"

1. Forward direction

stepwise\_base\_model\_fw <- stepAIC(trans\_model\_all, direction = "forward")

## Start: AIC=-950.7  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb +   
## pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP

par(mfrow = c(2, 2))  
 plot(stepwise\_base\_model\_fw)



summary(stepwise\_base\_model\_fw)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_h + batting\_2b + batting\_3b +   
## batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_hr +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP, data = transformed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.9324 -0.5430 -0.0133 0.5176 4.0620   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.343e-13 1.694e-02 0.000 1.000000   
## batting\_h -4.445e-02 1.009e-01 -0.441 0.659536   
## batting\_2b -4.135e-02 4.174e-02 -0.991 0.321924   
## batting\_3b 1.657e-01 3.972e-02 4.171 3.14e-05 \*\*\*  
## batting\_bb 1.950e-01 5.587e-02 3.490 0.000492 \*\*\*  
## batting\_so -2.226e-01 8.274e-02 -2.690 0.007200 \*\*   
## baserun\_sb 1.882e-01 4.701e-02 4.002 6.48e-05 \*\*\*  
## baserun\_cs 1.472e-01 4.353e-02 3.382 0.000733 \*\*\*  
## pitching\_hr 1.491e-02 7.246e-02 0.206 0.836935   
## pitching\_bb -1.950e-01 4.696e-02 -4.152 3.42e-05 \*\*\*  
## pitching\_so -9.011e-02 7.167e-02 -1.257 0.208772   
## fielding\_e -5.770e-01 3.884e-02 -14.858 < 2e-16 \*\*\*  
## fielding\_dp -1.951e-01 2.375e-02 -8.216 3.52e-16 \*\*\*  
## batting\_hbp\_bi -1.375e-01 2.145e-02 -6.413 1.73e-10 \*\*\*  
## batting\_1B 4.819e-02 7.787e-02 0.619 0.536041   
## total\_bases 3.530e-01 8.664e-02 4.075 4.77e-05 \*\*\*  
## total\_bases\_allowed 2.413e-01 5.831e-02 4.138 3.64e-05 \*\*\*  
## HR\_over\_OP -5.382e-02 3.604e-02 -1.493 0.135467   
## walks\_over\_OP 2.077e-01 5.003e-02 4.151 3.43e-05 \*\*\*  
## SO\_over\_OP 6.148e-02 3.833e-02 1.604 0.108868   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.808 on 2256 degrees of freedom  
## Multiple R-squared: 0.3526, Adjusted R-squared: 0.3472   
## F-statistic: 64.68 on 19 and 2256 DF, p-value: < 2.2e-16

paste('MSE equal ', mse(stepwise\_base\_model\_fw))

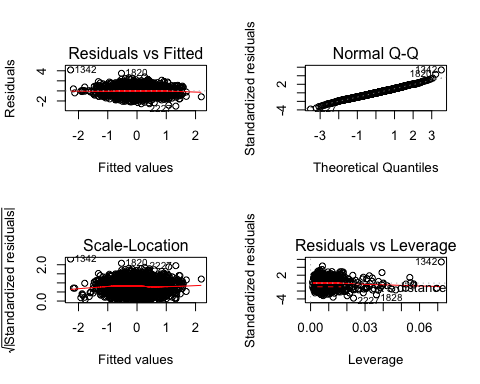
## [1] "MSE equal 0.647083867565306"

1. Backwards direction

stepwise\_base\_model\_bw <- stepAIC(trans\_model\_all, direction = "backward")

## Start: AIC=-950.7  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb +   
## pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - pitching\_hr 1 0.028 1472.8 -952.65  
## - batting\_h 1 0.127 1472.9 -952.50  
## - batting\_1B 1 0.250 1473.0 -952.31  
## - batting\_2b 1 0.641 1473.4 -951.71  
## - pitching\_so 1 1.032 1473.8 -951.10  
## <none> 1472.8 -950.70  
## - HR\_over\_OP 1 1.456 1474.2 -950.45  
## - SO\_over\_OP 1 1.679 1474.4 -950.10  
## - batting\_so 1 4.724 1477.5 -945.41  
## - baserun\_cs 1 7.465 1480.2 -941.19  
## - batting\_bb 1 7.952 1480.7 -940.44  
## - baserun\_sb 1 10.456 1483.2 -936.59  
## - total\_bases 1 10.838 1483.6 -936.01  
## - total\_bases\_allowed 1 11.176 1483.9 -935.49  
## - walks\_over\_OP 1 11.250 1484.0 -935.38  
## - pitching\_bb 1 11.255 1484.0 -935.37  
## - batting\_3b 1 11.358 1484.1 -935.21  
## - batting\_hbp\_bi 1 26.848 1499.6 -911.58  
## - fielding\_dp 1 44.062 1516.8 -885.60  
## - fielding\_e 1 144.108 1616.9 -740.23  
##   
## Step: AIC=-952.65  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_bb + pitching\_so +   
## fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B +   
## total\_bases + total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP +   
## SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - batting\_h 1 0.114 1472.9 -954.48  
## - batting\_1B 1 0.223 1473.0 -954.31  
## - pitching\_so 1 1.016 1473.8 -953.08  
## - batting\_2b 1 1.073 1473.9 -953.00  
## <none> 1472.8 -952.65  
## - SO\_over\_OP 1 1.684 1474.5 -952.05  
## - HR\_over\_OP 1 1.935 1474.7 -951.66  
## - batting\_so 1 4.696 1477.5 -947.41  
## - baserun\_cs 1 7.640 1480.4 -942.88  
## - batting\_bb 1 8.430 1481.2 -941.66  
## - pitching\_bb 1 11.290 1484.1 -937.27  
## - total\_bases\_allowed 1 11.433 1484.2 -937.05  
## - walks\_over\_OP 1 11.596 1484.4 -936.80  
## - baserun\_sb 1 12.371 1485.2 -935.61  
## - batting\_3b 1 13.795 1486.6 -933.43  
## - total\_bases 1 20.707 1493.5 -922.88  
## - batting\_hbp\_bi 1 28.450 1501.2 -911.11  
## - fielding\_dp 1 44.235 1517.0 -887.30  
## - fielding\_e 1 149.796 1622.6 -734.19  
##   
## Step: AIC=-954.48  
## target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + baserun\_cs + pitching\_bb + pitching\_so + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + batting\_1B + total\_bases +   
## total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP + SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - batting\_1B 1 0.151 1473.1 -956.24  
## - pitching\_so 1 0.945 1473.8 -955.02  
## <none> 1472.9 -954.48  
## - SO\_over\_OP 1 1.713 1474.6 -953.83  
## - HR\_over\_OP 1 1.999 1474.9 -953.39  
## - batting\_2b 1 2.643 1475.5 -952.40  
## - batting\_so 1 5.038 1477.9 -948.70  
## - baserun\_cs 1 7.574 1480.5 -944.80  
## - batting\_bb 1 9.108 1482.0 -942.44  
## - pitching\_bb 1 11.207 1484.1 -939.22  
## - walks\_over\_OP 1 11.516 1484.4 -938.75  
## - total\_bases\_allowed 1 13.437 1486.3 -935.81  
## - baserun\_sb 1 13.967 1486.9 -935.00  
## - batting\_3b 1 14.458 1487.4 -934.24  
## - total\_bases 1 26.309 1499.2 -916.18  
## - batting\_hbp\_bi 1 28.527 1501.4 -912.82  
## - fielding\_dp 1 44.141 1517.0 -889.27  
## - fielding\_e 1 149.682 1622.6 -736.19  
##   
## Step: AIC=-956.24  
## target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + baserun\_cs + pitching\_bb + pitching\_so + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP + walks\_over\_OP + SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - pitching\_so 1 0.982 1474.0 -956.73  
## <none> 1473.1 -956.24  
## - SO\_over\_OP 1 1.794 1474.8 -955.47  
## - HR\_over\_OP 1 1.933 1475.0 -955.26  
## - batting\_2b 1 2.677 1475.7 -954.11  
## - batting\_so 1 5.737 1478.8 -949.40  
## - baserun\_cs 1 7.950 1481.0 -945.99  
## - batting\_bb 1 8.967 1482.0 -944.43  
## - pitching\_bb 1 11.376 1484.4 -940.73  
## - walks\_over\_OP 1 11.520 1484.6 -940.51  
## - total\_bases\_allowed 1 13.510 1486.6 -937.46  
## - baserun\_sb 1 13.917 1487.0 -936.84  
## - batting\_3b 1 14.474 1487.5 -935.99  
## - total\_bases 1 29.324 1502.4 -913.38  
## - batting\_hbp\_bi 1 29.810 1502.9 -912.64  
## - fielding\_dp 1 44.052 1517.1 -891.18  
## - fielding\_e 1 149.533 1622.6 -738.19  
##   
## Step: AIC=-956.73  
## target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - SO\_over\_OP 1 1.218 1475.3 -956.85  
## <none> 1474.0 -956.73  
## - HR\_over\_OP 1 1.866 1475.9 -955.85  
## - batting\_2b 1 2.989 1477.0 -954.12  
## - baserun\_cs 1 7.544 1481.6 -947.11  
## - batting\_bb 1 9.823 1483.9 -943.61  
## - walks\_over\_OP 1 11.345 1485.4 -941.28  
## - pitching\_bb 1 12.096 1486.1 -940.13  
## - baserun\_sb 1 13.390 1487.4 -938.14  
## - total\_bases\_allowed 1 13.685 1487.7 -937.69  
## - batting\_3b 1 14.198 1488.2 -936.91  
## - batting\_hbp\_bi 1 30.730 1504.8 -911.77  
## - total\_bases 1 40.213 1514.2 -897.47  
## - fielding\_dp 1 43.282 1517.3 -892.86  
## - batting\_so 1 73.325 1547.4 -848.23  
## - fielding\_e 1 148.828 1622.9 -739.80  
##   
## Step: AIC=-956.85  
## target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## <none> 1475.3 -956.85  
## - batting\_2b 1 2.697 1478.0 -954.69  
## - HR\_over\_OP 1 3.776 1479.0 -953.03  
## - baserun\_cs 1 8.833 1484.1 -945.26  
## - batting\_bb 1 9.329 1484.6 -944.50  
## - walks\_over\_OP 1 10.136 1485.4 -943.26  
## - pitching\_bb 1 11.722 1487.0 -940.83  
## - baserun\_sb 1 13.569 1488.8 -938.01  
## - batting\_3b 1 14.061 1489.3 -937.26  
## - total\_bases\_allowed 1 14.737 1490.0 -936.22  
## - batting\_hbp\_bi 1 30.884 1506.1 -911.69  
## - total\_bases 1 38.995 1514.2 -899.47  
## - fielding\_dp 1 43.296 1518.5 -893.01  
## - batting\_so 1 74.474 1549.7 -846.75  
## - fielding\_e 1 148.669 1623.9 -740.32

par(mfrow = c(2, 2))  
 plot(stepwise\_base\_model\_bw)



summary(stepwise\_base\_model\_bw)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP + walks\_over\_OP, data = transformed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.0325 -0.5448 -0.0108 0.5230 4.1704   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.764e-15 1.693e-02 0.000 1.000000   
## batting\_2b -5.547e-02 2.728e-02 -2.033 0.042159 \*   
## batting\_3b 1.549e-01 3.337e-02 4.642 3.64e-06 \*\*\*  
## batting\_bb 1.950e-01 5.158e-02 3.781 0.000160 \*\*\*  
## batting\_so -3.124e-01 2.924e-02 -10.684 < 2e-16 \*\*\*  
## baserun\_sb 1.851e-01 4.059e-02 4.560 5.38e-06 \*\*\*  
## baserun\_cs 1.552e-01 4.219e-02 3.679 0.000239 \*\*\*  
## pitching\_bb -1.952e-01 4.605e-02 -4.239 2.34e-05 \*\*\*  
## fielding\_e -5.754e-01 3.812e-02 -15.095 < 2e-16 \*\*\*  
## fielding\_dp -1.910e-01 2.345e-02 -8.146 6.15e-16 \*\*\*  
## batting\_hbp\_bi -1.423e-01 2.068e-02 -6.880 7.72e-12 \*\*\*  
## total\_bases 3.721e-01 4.813e-02 7.731 1.60e-14 \*\*\*  
## total\_bases\_allowed 2.054e-01 4.323e-02 4.752 2.14e-06 \*\*\*  
## HR\_over\_OP -7.286e-02 3.029e-02 -2.406 0.016226 \*   
## walks\_over\_OP 1.805e-01 4.580e-02 3.941 8.35e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8078 on 2261 degrees of freedom  
## Multiple R-squared: 0.3515, Adjusted R-squared: 0.3475   
## F-statistic: 87.55 on 14 and 2261 DF, p-value: < 2.2e-16

paste('MSE equal ', mse(stepwise\_base\_model\_bw))

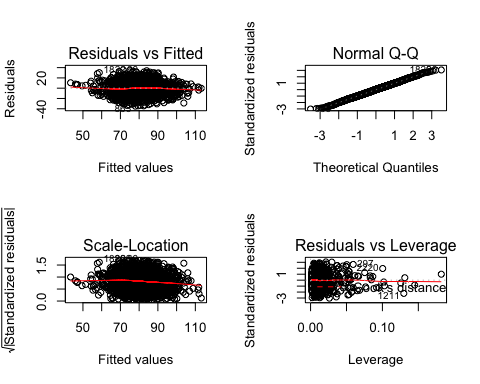
## [1] "MSE equal 0.648179170553685"

Let's remove influential observations based on cook's distance chart. We will remove the following observations: 1342, 1810, 1828, 2136, 1820, 2227,1340, 1811, 2233, 1896, 2020, 2228.

Those observations will be removed from these datasets: transformed and outlier\_treat

outlier\_treat\_rm <- outlier\_treat[-c(1342, 1810, 1828, 2136, 1820, 2227,1340, 1811, 2233, 1896, 2020, 2228),]  
transformed\_rm <- transformed[-c(1342, 1810, 1828, 2136, 1820, 2227,1340, 1811, 2233, 1896, 2020, 2228),]

base\_model\_orig\_rm <-  
 lm(target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP + SO\_over\_OP, data = outlier\_treat\_rm)  
 par(mfrow = c(2, 2))  
 plot(base\_model\_orig\_rm)



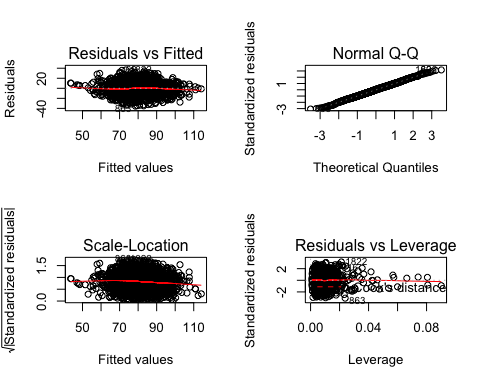
summary(base\_model\_orig\_rm)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_h + batting\_2b + batting\_3b +   
## batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_hr +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP, data = outlier\_treat\_rm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -34.935 -7.981 0.192 7.807 35.524   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34.3291551 6.0483138 5.676 1.56e-08 \*\*\*  
## batting\_h 0.0218895 0.0136362 1.605 0.108580   
## batting\_2b -0.0627070 0.0152440 -4.114 4.04e-05 \*\*\*  
## batting\_3b 0.0405900 0.0230262 1.763 0.078074 .   
## batting\_bb 0.0411676 0.0097664 4.215 2.59e-05 \*\*\*  
## batting\_so -0.0160349 0.0056546 -2.836 0.004613 \*\*   
## baserun\_sb 0.0605082 0.0089420 6.767 1.67e-11 \*\*\*  
## baserun\_cs -0.0039357 0.0172741 -0.228 0.819795   
## pitching\_hr -0.0410934 0.0200349 -2.051 0.040374 \*   
## pitching\_bb -0.0502903 0.0081714 -6.154 8.90e-10 \*\*\*  
## pitching\_so -0.0009172 0.0052280 -0.175 0.860747   
## fielding\_e -0.0441196 0.0030316 -14.553 < 2e-16 \*\*\*  
## fielding\_dp -0.0993091 0.0129565 -7.665 2.65e-14 \*\*\*  
## batting\_hbp\_bi -4.2752572 1.1134687 -3.840 0.000127 \*\*\*  
## batting\_1B -0.0287308 0.0133420 -2.153 0.031392 \*   
## total\_bases 0.0248649 0.0047017 5.288 1.35e-07 \*\*\*  
## total\_bases\_allowed 0.0117947 0.0021119 5.585 2.62e-08 \*\*\*  
## HR\_over\_OP -0.1019108 0.0816525 -1.248 0.212123   
## walks\_over\_OP 0.0266495 0.0108719 2.451 0.014313 \*   
## SO\_over\_OP 0.0091144 0.0043901 2.076 0.037994 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.71 on 2244 degrees of freedom  
## Multiple R-squared: 0.3642, Adjusted R-squared: 0.3588   
## F-statistic: 67.64 on 19 and 2244 DF, p-value: < 2.2e-16

paste('MSE equal ', mse(base\_model\_orig\_rm))

## [1] "MSE equal 136.021498431333"

base\_model\_lp\_rm <-  
 lm(target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so + baserun\_sb + pitching\_bb +   
 fielding\_e + fielding\_dp + batting\_hbp\_bi + total\_bases + total\_bases\_allowed + walks\_over\_OP + SO\_over\_OP, data = outlier\_treat\_rm)  
 par(mfrow = c(2, 2))  
 plot(base\_model\_lp\_rm)



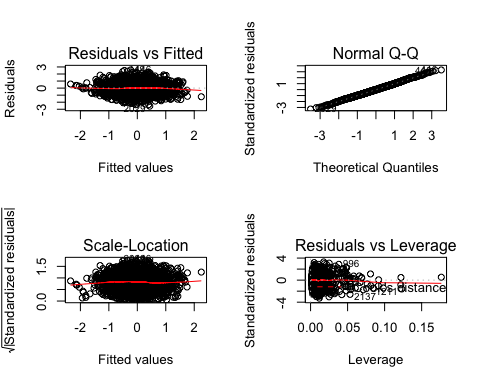
summary(base\_model\_lp\_rm)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + pitching\_bb + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + total\_bases + total\_bases\_allowed + walks\_over\_OP +   
## SO\_over\_OP, data = outlier\_treat\_rm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -35.554 -7.946 0.194 7.746 36.628   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.423369 3.214253 9.776 < 2e-16 \*\*\*  
## batting\_2b -0.034378 0.008980 -3.828 0.000133 \*\*\*  
## batting\_3b 0.071554 0.017427 4.106 4.17e-05 \*\*\*  
## batting\_bb 0.045871 0.009117 5.031 5.26e-07 \*\*\*  
## batting\_so -0.016883 0.001772 -9.528 < 2e-16 \*\*\*  
## baserun\_sb 0.062321 0.005752 10.834 < 2e-16 \*\*\*  
## pitching\_bb -0.050756 0.008068 -6.291 3.78e-10 \*\*\*  
## fielding\_e -0.042639 0.002883 -14.787 < 2e-16 \*\*\*  
## fielding\_dp -0.100316 0.012606 -7.958 2.75e-15 \*\*\*  
## batting\_hbp\_bi -3.980375 1.069020 -3.723 0.000201 \*\*\*  
## total\_bases 0.020573 0.002454 8.382 < 2e-16 \*\*\*  
## total\_bases\_allowed 0.012093 0.001508 8.018 1.71e-15 \*\*\*  
## walks\_over\_OP 0.027087 0.010490 2.582 0.009879 \*\*   
## SO\_over\_OP 0.009524 0.003845 2.477 0.013320 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.72 on 2250 degrees of freedom  
## Multiple R-squared: 0.3623, Adjusted R-squared: 0.3586   
## F-statistic: 98.33 on 13 and 2250 DF, p-value: < 2.2e-16

paste('MSE equal ', mse(base\_model\_lp\_rm))

## [1] "MSE equal 136.418202030273"

trans\_model\_all\_rm <-  
 lm(target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP + SO\_over\_OP, data = transformed\_rm)  
 par(mfrow = c(2, 2))  
 plot(trans\_model\_all\_rm)



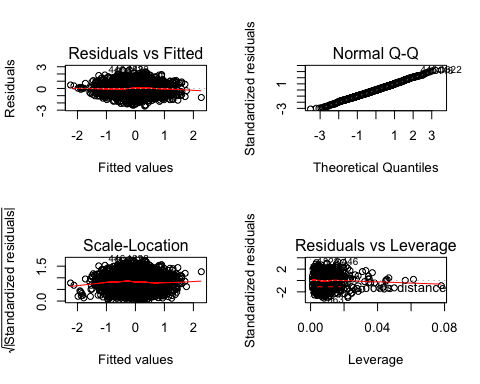
summary(trans\_model\_all\_rm)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_h + batting\_2b + batting\_3b +   
## batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_hr +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP, data = transformed\_rm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.56019 -0.54095 -0.00762 0.51300 2.57939   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.004208 0.016568 -0.254 0.799545   
## batting\_h 0.074177 0.100439 0.739 0.460270   
## batting\_2b -0.109759 0.043200 -2.541 0.011130 \*   
## batting\_3b 0.150447 0.039975 3.764 0.000172 \*\*\*  
## batting\_bb 0.292214 0.058825 4.968 7.29e-07 \*\*\*  
## batting\_so -0.310532 0.087329 -3.556 0.000384 \*\*\*  
## baserun\_sb 0.172427 0.047514 3.629 0.000291 \*\*\*  
## baserun\_cs 0.135214 0.043083 3.138 0.001720 \*\*   
## pitching\_hr -0.054814 0.078107 -0.702 0.482888   
## pitching\_bb -0.283055 0.048945 -5.783 8.36e-09 \*\*\*  
## pitching\_so 0.001335 0.075176 0.018 0.985830   
## fielding\_e -0.595271 0.038211 -15.578 < 2e-16 \*\*\*  
## fielding\_dp -0.194105 0.023177 -8.375 < 2e-16 \*\*\*  
## batting\_hbp\_bi -0.137895 0.021106 -6.534 7.92e-11 \*\*\*  
## batting\_1B -0.043597 0.079796 -0.546 0.584874   
## total\_bases 0.405897 0.089732 4.523 6.40e-06 \*\*\*  
## total\_bases\_allowed 0.175163 0.059663 2.936 0.003360 \*\*   
## HR\_over\_OP -0.104410 0.035925 -2.906 0.003693 \*\*   
## walks\_over\_OP 0.174683 0.049432 3.534 0.000418 \*\*\*  
## SO\_over\_OP 0.031538 0.037757 0.835 0.403647   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7881 on 2244 degrees of freedom  
## Multiple R-squared: 0.3709, Adjusted R-squared: 0.3656   
## F-statistic: 69.64 on 19 and 2244 DF, p-value: < 2.2e-16

paste('MSE equal ', mse(trans\_model\_all\_rm))

## [1] "MSE equal 0.615551935849093"

trans\_model\_lp\_rm <-  
 lm(target\_wins ~ batting\_3b + batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e + fielding\_dp + batting\_hbp\_bi + total\_bases + total\_bases\_allowed + walks\_over\_OP + SO\_over\_OP, data = transformed\_rm)  
 par(mfrow = c(2, 2))  
 plot(trans\_model\_lp\_rm)



summary(trans\_model\_lp\_rm)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + total\_bases + total\_bases\_allowed + walks\_over\_OP +   
## SO\_over\_OP, data = transformed\_rm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.44678 -0.54976 -0.00499 0.51233 2.56018   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.003734 0.016609 -0.225 0.82213   
## batting\_3b 0.193489 0.032339 5.983 2.54e-09 \*\*\*  
## batting\_bb 0.305651 0.053250 5.740 1.08e-08 \*\*\*  
## batting\_so -0.318539 0.029906 -10.651 < 2e-16 \*\*\*  
## baserun\_sb 0.190847 0.039415 4.842 1.37e-06 \*\*\*  
## baserun\_cs 0.126372 0.041996 3.009 0.00265 \*\*   
## pitching\_bb -0.275581 0.048522 -5.680 1.53e-08 \*\*\*  
## fielding\_e -0.579511 0.037258 -15.554 < 2e-16 \*\*\*  
## fielding\_dp -0.192350 0.022915 -8.394 < 2e-16 \*\*\*  
## batting\_hbp\_bi -0.152738 0.019816 -7.708 1.91e-14 \*\*\*  
## total\_bases 0.320545 0.041004 7.817 8.21e-15 \*\*\*  
## total\_bases\_allowed 0.223549 0.042162 5.302 1.26e-07 \*\*\*  
## walks\_over\_OP 0.156725 0.047703 3.285 0.00103 \*\*   
## SO\_over\_OP 0.060558 0.033262 1.821 0.06879 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7901 on 2250 degrees of freedom  
## Multiple R-squared: 0.366, Adjusted R-squared: 0.3623   
## F-statistic: 99.9 on 13 and 2250 DF, p-value: < 2.2e-16

paste('MSE equal ', mse(trans\_model\_lp\_rm))

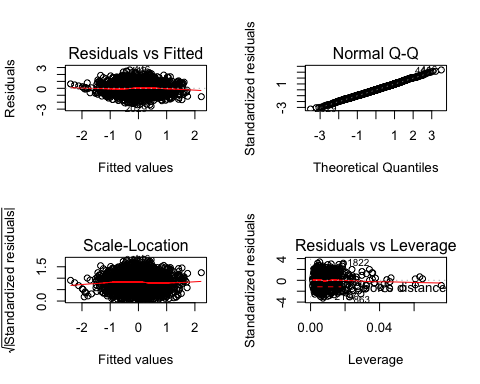
## [1] "MSE equal 0.620413235273429"

1. Stepwise Both direction

stepwise\_base\_model\_bd\_rm <- stepAIC(trans\_model\_all\_rm, direction = "both")

## Start: AIC=-1058.57  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb +   
## pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - pitching\_so 1 0.000 1393.6 -1060.57  
## - batting\_1B 1 0.185 1393.8 -1060.27  
## - pitching\_hr 1 0.306 1393.9 -1060.08  
## - batting\_h 1 0.339 1394.0 -1060.02  
## - SO\_over\_OP 1 0.433 1394.0 -1059.87  
## <none> 1393.6 -1058.57  
## - batting\_2b 1 4.009 1397.6 -1054.07  
## - HR\_over\_OP 1 5.246 1398.9 -1052.07  
## - total\_bases\_allowed 1 5.353 1399.0 -1051.89  
## - baserun\_cs 1 6.117 1399.7 -1050.66  
## - walks\_over\_OP 1 7.755 1401.4 -1048.01  
## - batting\_so 1 7.853 1401.5 -1047.85  
## - baserun\_sb 1 8.179 1401.8 -1047.33  
## - batting\_3b 1 8.796 1402.4 -1046.33  
## - total\_bases 1 12.707 1406.3 -1040.02  
## - batting\_bb 1 15.325 1408.9 -1035.81  
## - pitching\_bb 1 20.770 1414.4 -1027.08  
## - batting\_hbp\_bi 1 26.510 1420.1 -1017.91  
## - fielding\_dp 1 43.560 1437.2 -990.89  
## - fielding\_e 1 150.717 1544.3 -828.08  
##   
## Step: AIC=-1060.57  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb +   
## fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B +   
## total\_bases + total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP +   
## SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - batting\_1B 1 0.186 1393.8 -1062.27  
## - pitching\_hr 1 0.306 1393.9 -1062.08  
## - batting\_h 1 0.342 1394.0 -1062.02  
## - SO\_over\_OP 1 0.480 1394.1 -1061.79  
## <none> 1393.6 -1060.57  
## + pitching\_so 1 0.000 1393.6 -1058.57  
## - batting\_2b 1 4.085 1397.7 -1055.95  
## - HR\_over\_OP 1 5.246 1398.9 -1054.07  
## - baserun\_cs 1 6.217 1399.8 -1052.50  
## - total\_bases\_allowed 1 7.749 1401.4 -1050.02  
## - walks\_over\_OP 1 7.764 1401.4 -1050.00  
## - baserun\_sb 1 8.192 1401.8 -1049.30  
## - batting\_3b 1 8.847 1402.5 -1048.25  
## - total\_bases 1 13.290 1406.9 -1041.09  
## - batting\_bb 1 15.460 1409.1 -1037.60  
## - pitching\_bb 1 20.790 1414.4 -1029.05  
## - batting\_hbp\_bi 1 26.595 1420.2 -1019.78  
## - batting\_so 1 38.881 1432.5 -1000.27  
## - fielding\_dp 1 43.798 1437.4 -992.52  
## - fielding\_e 1 150.878 1544.5 -829.85  
##   
## Step: AIC=-1062.27  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb +   
## fielding\_e + fielding\_dp + batting\_hbp\_bi + total\_bases +   
## total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP + SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - pitching\_hr 1 0.168 1394.0 -1064.00  
## - batting\_h 1 0.184 1394.0 -1063.97  
## - SO\_over\_OP 1 0.437 1394.2 -1063.56  
## <none> 1393.8 -1062.27  
## + batting\_1B 1 0.186 1393.6 -1060.57  
## + pitching\_so 1 0.001 1393.8 -1060.27  
## - HR\_over\_OP 1 5.061 1398.9 -1056.07  
## - batting\_2b 1 5.513 1399.3 -1055.33  
## - baserun\_cs 1 6.085 1399.9 -1054.41  
## - walks\_over\_OP 1 8.071 1401.9 -1051.20  
## - baserun\_sb 1 8.242 1402.0 -1050.92  
## - total\_bases\_allowed 1 9.351 1403.2 -1049.13  
## - batting\_3b 1 12.174 1406.0 -1044.58  
## - total\_bases 1 13.232 1407.0 -1042.88  
## - batting\_bb 1 15.519 1409.3 -1039.20  
## - pitching\_bb 1 21.116 1414.9 -1030.23  
## - batting\_hbp\_bi 1 26.603 1420.4 -1021.47  
## - batting\_so 1 39.044 1432.8 -1001.72  
## - fielding\_dp 1 44.707 1438.5 -992.79  
## - fielding\_e 1 150.759 1544.5 -831.75  
##   
## Step: AIC=-1064  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP + walks\_over\_OP + SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - SO\_over\_OP 1 0.446 1394.4 -1065.28  
## - batting\_h 1 0.517 1394.5 -1065.16  
## <none> 1394.0 -1064.00  
## + pitching\_hr 1 0.168 1393.8 -1062.27  
## + batting\_1B 1 0.048 1393.9 -1062.08  
## + pitching\_so 1 0.000 1394.0 -1062.00  
## - HR\_over\_OP 1 5.095 1399.1 -1057.74  
## - batting\_2b 1 5.759 1399.7 -1056.66  
## - baserun\_cs 1 5.929 1399.9 -1056.39  
## - walks\_over\_OP 1 7.906 1401.9 -1053.19  
## - total\_bases\_allowed 1 9.668 1403.6 -1050.35  
## - baserun\_sb 1 13.204 1407.2 -1044.65  
## - batting\_3b 1 15.901 1409.9 -1040.32  
## - batting\_bb 1 18.722 1412.7 -1035.79  
## - pitching\_bb 1 20.956 1414.9 -1032.22  
## - total\_bases 1 26.226 1420.2 -1023.80  
## - batting\_hbp\_bi 1 26.772 1420.7 -1022.93  
## - fielding\_dp 1 45.200 1439.2 -993.75  
## - batting\_so 1 49.132 1443.1 -987.58  
## - fielding\_e 1 154.706 1548.7 -827.72  
##   
## Step: AIC=-1065.28  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP + walks\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - batting\_h 1 0.537 1395.0 -1066.40  
## <none> 1394.4 -1065.28  
## + SO\_over\_OP 1 0.446 1394.0 -1064.00  
## + pitching\_hr 1 0.177 1394.2 -1063.56  
## + pitching\_so 1 0.038 1394.4 -1063.34  
## + batting\_1B 1 0.028 1394.4 -1063.32  
## - batting\_2b 1 5.559 1400.0 -1058.27  
## - baserun\_cs 1 6.655 1401.1 -1056.50  
## - HR\_over\_OP 1 7.344 1401.8 -1055.38  
## - walks\_over\_OP 1 7.590 1402.0 -1054.98  
## - total\_bases\_allowed 1 10.220 1404.6 -1050.74  
## - baserun\_sb 1 13.317 1407.7 -1045.76  
## - batting\_3b 1 15.810 1410.2 -1041.75  
## - batting\_bb 1 18.469 1412.9 -1037.49  
## - pitching\_bb 1 20.773 1415.2 -1033.80  
## - total\_bases 1 25.813 1420.2 -1025.75  
## - batting\_hbp\_bi 1 26.829 1421.2 -1024.13  
## - fielding\_dp 1 45.226 1439.6 -995.01  
## - batting\_so 1 50.248 1444.7 -987.13  
## - fielding\_e 1 154.618 1549.0 -829.20  
##   
## Step: AIC=-1066.4  
## target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## <none> 1395.0 -1066.40  
## + batting\_h 1 0.537 1394.4 -1065.28  
## + pitching\_hr 1 0.524 1394.4 -1065.25  
## + SO\_over\_OP 1 0.466 1394.5 -1065.16  
## + batting\_1B 1 0.349 1394.6 -1064.97  
## + pitching\_so 1 0.017 1394.9 -1064.43  
## - batting\_2b 1 5.022 1400.0 -1060.27  
## - HR\_over\_OP 1 7.003 1402.0 -1057.07  
## - baserun\_cs 1 7.244 1402.2 -1056.68  
## - walks\_over\_OP 1 7.900 1402.8 -1055.62  
## - total\_bases\_allowed 1 12.071 1407.0 -1048.90  
## - baserun\_sb 1 12.826 1407.8 -1047.68  
## - batting\_3b 1 16.921 1411.9 -1041.11  
## - batting\_bb 1 17.937 1412.9 -1039.48  
## - pitching\_bb 1 21.414 1416.4 -1033.91  
## - batting\_hbp\_bi 1 28.433 1423.4 -1022.72  
## - total\_bases 1 40.741 1435.7 -1003.23  
## - fielding\_dp 1 44.731 1439.7 -996.94  
## - batting\_so 1 78.348 1473.3 -944.69  
## - fielding\_e 1 154.269 1549.2 -830.93

par(mfrow = c(2, 2))  
 plot(stepwise\_base\_model\_bd\_rm)



summary(stepwise\_base\_model\_bd\_rm)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP + walks\_over\_OP, data = transformed\_rm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.5712 -0.5427 -0.0066 0.5141 2.6560   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.004106 0.016556 -0.248 0.804132   
## batting\_2b -0.076989 0.027057 -2.845 0.004476 \*\*   
## batting\_3b 0.170960 0.032731 5.223 1.92e-07 \*\*\*  
## batting\_bb 0.290798 0.054075 5.378 8.33e-08 \*\*\*  
## batting\_so -0.322686 0.028711 -11.239 < 2e-16 \*\*\*  
## baserun\_sb 0.181398 0.039890 4.547 5.72e-06 \*\*\*  
## baserun\_cs 0.141448 0.041390 3.417 0.000643 \*\*\*  
## pitching\_bb -0.284971 0.048500 -5.876 4.84e-09 \*\*\*  
## fielding\_e -0.589229 0.037362 -15.771 < 2e-16 \*\*\*  
## fielding\_dp -0.194402 0.022892 -8.492 < 2e-16 \*\*\*  
## batting\_hbp\_bi -0.136648 0.020182 -6.771 1.63e-11 \*\*\*  
## total\_bases 0.388993 0.047997 8.105 8.59e-16 \*\*\*  
## total\_bases\_allowed 0.189220 0.042893 4.411 1.08e-05 \*\*\*  
## HR\_over\_OP -0.101620 0.030242 -3.360 0.000792 \*\*\*  
## walks\_over\_OP 0.161277 0.045189 3.569 0.000366 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7876 on 2249 degrees of freedom  
## Multiple R-squared: 0.3703, Adjusted R-squared: 0.3664   
## F-statistic: 94.48 on 14 and 2249 DF, p-value: < 2.2e-16

paste('MSE equal ', mse(stepwise\_base\_model\_bd\_rm))

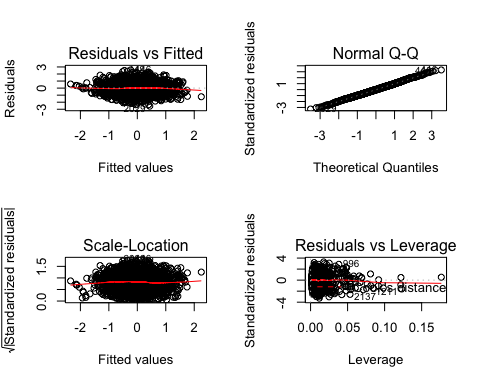
## [1] "MSE equal 0.616142519793743"

1. Forward direction

stepwise\_base\_model\_fw\_rm <- stepAIC(trans\_model\_all\_rm, direction = "forward")

## Start: AIC=-1058.57  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb +   
## pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP

par(mfrow = c(2, 2))  
 plot(stepwise\_base\_model\_fw\_rm)



summary(stepwise\_base\_model\_fw\_rm)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_h + batting\_2b + batting\_3b +   
## batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_hr +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP, data = transformed\_rm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.56019 -0.54095 -0.00762 0.51300 2.57939   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.004208 0.016568 -0.254 0.799545   
## batting\_h 0.074177 0.100439 0.739 0.460270   
## batting\_2b -0.109759 0.043200 -2.541 0.011130 \*   
## batting\_3b 0.150447 0.039975 3.764 0.000172 \*\*\*  
## batting\_bb 0.292214 0.058825 4.968 7.29e-07 \*\*\*  
## batting\_so -0.310532 0.087329 -3.556 0.000384 \*\*\*  
## baserun\_sb 0.172427 0.047514 3.629 0.000291 \*\*\*  
## baserun\_cs 0.135214 0.043083 3.138 0.001720 \*\*   
## pitching\_hr -0.054814 0.078107 -0.702 0.482888   
## pitching\_bb -0.283055 0.048945 -5.783 8.36e-09 \*\*\*  
## pitching\_so 0.001335 0.075176 0.018 0.985830   
## fielding\_e -0.595271 0.038211 -15.578 < 2e-16 \*\*\*  
## fielding\_dp -0.194105 0.023177 -8.375 < 2e-16 \*\*\*  
## batting\_hbp\_bi -0.137895 0.021106 -6.534 7.92e-11 \*\*\*  
## batting\_1B -0.043597 0.079796 -0.546 0.584874   
## total\_bases 0.405897 0.089732 4.523 6.40e-06 \*\*\*  
## total\_bases\_allowed 0.175163 0.059663 2.936 0.003360 \*\*   
## HR\_over\_OP -0.104410 0.035925 -2.906 0.003693 \*\*   
## walks\_over\_OP 0.174683 0.049432 3.534 0.000418 \*\*\*  
## SO\_over\_OP 0.031538 0.037757 0.835 0.403647   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7881 on 2244 degrees of freedom  
## Multiple R-squared: 0.3709, Adjusted R-squared: 0.3656   
## F-statistic: 69.64 on 19 and 2244 DF, p-value: < 2.2e-16

paste('MSE equal ', mse(stepwise\_base\_model\_fw\_rm))

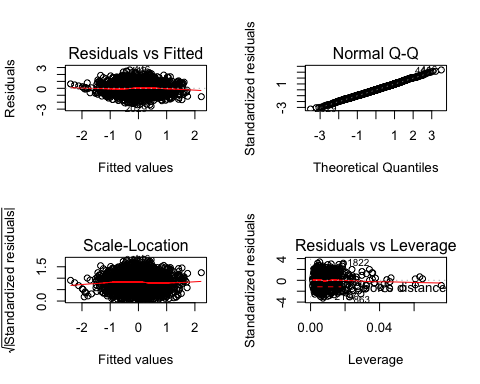
## [1] "MSE equal 0.615551935849093"

1. Backwards direction

stepwise\_base\_model\_bw\_rm <- stepAIC(trans\_model\_all\_rm, direction = "backward")

## Start: AIC=-1058.57  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb +   
## pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - pitching\_so 1 0.000 1393.6 -1060.57  
## - batting\_1B 1 0.185 1393.8 -1060.27  
## - pitching\_hr 1 0.306 1393.9 -1060.08  
## - batting\_h 1 0.339 1394.0 -1060.02  
## - SO\_over\_OP 1 0.433 1394.0 -1059.87  
## <none> 1393.6 -1058.57  
## - batting\_2b 1 4.009 1397.6 -1054.07  
## - HR\_over\_OP 1 5.246 1398.9 -1052.07  
## - total\_bases\_allowed 1 5.353 1399.0 -1051.89  
## - baserun\_cs 1 6.117 1399.7 -1050.66  
## - walks\_over\_OP 1 7.755 1401.4 -1048.01  
## - batting\_so 1 7.853 1401.5 -1047.85  
## - baserun\_sb 1 8.179 1401.8 -1047.33  
## - batting\_3b 1 8.796 1402.4 -1046.33  
## - total\_bases 1 12.707 1406.3 -1040.02  
## - batting\_bb 1 15.325 1408.9 -1035.81  
## - pitching\_bb 1 20.770 1414.4 -1027.08  
## - batting\_hbp\_bi 1 26.510 1420.1 -1017.91  
## - fielding\_dp 1 43.560 1437.2 -990.89  
## - fielding\_e 1 150.717 1544.3 -828.08  
##   
## Step: AIC=-1060.57  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb +   
## fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B +   
## total\_bases + total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP +   
## SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - batting\_1B 1 0.186 1393.8 -1062.27  
## - pitching\_hr 1 0.306 1393.9 -1062.08  
## - batting\_h 1 0.342 1394.0 -1062.02  
## - SO\_over\_OP 1 0.480 1394.1 -1061.79  
## <none> 1393.6 -1060.57  
## - batting\_2b 1 4.085 1397.7 -1055.95  
## - HR\_over\_OP 1 5.246 1398.9 -1054.07  
## - baserun\_cs 1 6.217 1399.8 -1052.50  
## - total\_bases\_allowed 1 7.749 1401.4 -1050.02  
## - walks\_over\_OP 1 7.764 1401.4 -1050.00  
## - baserun\_sb 1 8.192 1401.8 -1049.30  
## - batting\_3b 1 8.847 1402.5 -1048.25  
## - total\_bases 1 13.290 1406.9 -1041.09  
## - batting\_bb 1 15.460 1409.1 -1037.60  
## - pitching\_bb 1 20.790 1414.4 -1029.05  
## - batting\_hbp\_bi 1 26.595 1420.2 -1019.78  
## - batting\_so 1 38.881 1432.5 -1000.27  
## - fielding\_dp 1 43.798 1437.4 -992.52  
## - fielding\_e 1 150.878 1544.5 -829.85  
##   
## Step: AIC=-1062.27  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_hr + pitching\_bb +   
## fielding\_e + fielding\_dp + batting\_hbp\_bi + total\_bases +   
## total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP + SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - pitching\_hr 1 0.168 1394.0 -1064.00  
## - batting\_h 1 0.184 1394.0 -1063.97  
## - SO\_over\_OP 1 0.437 1394.2 -1063.56  
## <none> 1393.8 -1062.27  
## - HR\_over\_OP 1 5.061 1398.9 -1056.07  
## - batting\_2b 1 5.513 1399.3 -1055.33  
## - baserun\_cs 1 6.085 1399.9 -1054.41  
## - walks\_over\_OP 1 8.071 1401.9 -1051.20  
## - baserun\_sb 1 8.242 1402.0 -1050.92  
## - total\_bases\_allowed 1 9.351 1403.2 -1049.13  
## - batting\_3b 1 12.174 1406.0 -1044.58  
## - total\_bases 1 13.232 1407.0 -1042.88  
## - batting\_bb 1 15.519 1409.3 -1039.20  
## - pitching\_bb 1 21.116 1414.9 -1030.23  
## - batting\_hbp\_bi 1 26.603 1420.4 -1021.47  
## - batting\_so 1 39.044 1432.8 -1001.72  
## - fielding\_dp 1 44.707 1438.5 -992.79  
## - fielding\_e 1 150.759 1544.5 -831.75  
##   
## Step: AIC=-1064  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP + walks\_over\_OP + SO\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - SO\_over\_OP 1 0.446 1394.4 -1065.28  
## - batting\_h 1 0.517 1394.5 -1065.16  
## <none> 1394.0 -1064.00  
## - HR\_over\_OP 1 5.095 1399.1 -1057.74  
## - batting\_2b 1 5.759 1399.7 -1056.66  
## - baserun\_cs 1 5.929 1399.9 -1056.39  
## - walks\_over\_OP 1 7.906 1401.9 -1053.19  
## - total\_bases\_allowed 1 9.668 1403.6 -1050.35  
## - baserun\_sb 1 13.204 1407.2 -1044.65  
## - batting\_3b 1 15.901 1409.9 -1040.32  
## - batting\_bb 1 18.722 1412.7 -1035.79  
## - pitching\_bb 1 20.956 1414.9 -1032.22  
## - total\_bases 1 26.226 1420.2 -1023.80  
## - batting\_hbp\_bi 1 26.772 1420.7 -1022.93  
## - fielding\_dp 1 45.200 1439.2 -993.75  
## - batting\_so 1 49.132 1443.1 -987.58  
## - fielding\_e 1 154.706 1548.7 -827.72  
##   
## Step: AIC=-1065.28  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP + walks\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - batting\_h 1 0.537 1395.0 -1066.40  
## <none> 1394.4 -1065.28  
## - batting\_2b 1 5.559 1400.0 -1058.27  
## - baserun\_cs 1 6.655 1401.1 -1056.50  
## - HR\_over\_OP 1 7.344 1401.8 -1055.38  
## - walks\_over\_OP 1 7.590 1402.0 -1054.98  
## - total\_bases\_allowed 1 10.220 1404.6 -1050.74  
## - baserun\_sb 1 13.317 1407.7 -1045.76  
## - batting\_3b 1 15.810 1410.2 -1041.75  
## - batting\_bb 1 18.469 1412.9 -1037.49  
## - pitching\_bb 1 20.773 1415.2 -1033.80  
## - total\_bases 1 25.813 1420.2 -1025.75  
## - batting\_hbp\_bi 1 26.829 1421.2 -1024.13  
## - fielding\_dp 1 45.226 1439.6 -995.01  
## - batting\_so 1 50.248 1444.7 -987.13  
## - fielding\_e 1 154.618 1549.0 -829.20  
##   
## Step: AIC=-1066.4  
## target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## <none> 1395.0 -1066.40  
## - batting\_2b 1 5.022 1400.0 -1060.27  
## - HR\_over\_OP 1 7.003 1402.0 -1057.07  
## - baserun\_cs 1 7.244 1402.2 -1056.68  
## - walks\_over\_OP 1 7.900 1402.8 -1055.62  
## - total\_bases\_allowed 1 12.071 1407.0 -1048.90  
## - baserun\_sb 1 12.826 1407.8 -1047.68  
## - batting\_3b 1 16.921 1411.9 -1041.11  
## - batting\_bb 1 17.937 1412.9 -1039.48  
## - pitching\_bb 1 21.414 1416.4 -1033.91  
## - batting\_hbp\_bi 1 28.433 1423.4 -1022.72  
## - total\_bases 1 40.741 1435.7 -1003.23  
## - fielding\_dp 1 44.731 1439.7 -996.94  
## - batting\_so 1 78.348 1473.3 -944.69  
## - fielding\_e 1 154.269 1549.2 -830.93

par(mfrow = c(2, 2))  
 plot(stepwise\_base\_model\_bw\_rm)



summary(stepwise\_base\_model\_bw\_rm)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_bb + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP + walks\_over\_OP, data = transformed\_rm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.5712 -0.5427 -0.0066 0.5141 2.6560   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.004106 0.016556 -0.248 0.804132   
## batting\_2b -0.076989 0.027057 -2.845 0.004476 \*\*   
## batting\_3b 0.170960 0.032731 5.223 1.92e-07 \*\*\*  
## batting\_bb 0.290798 0.054075 5.378 8.33e-08 \*\*\*  
## batting\_so -0.322686 0.028711 -11.239 < 2e-16 \*\*\*  
## baserun\_sb 0.181398 0.039890 4.547 5.72e-06 \*\*\*  
## baserun\_cs 0.141448 0.041390 3.417 0.000643 \*\*\*  
## pitching\_bb -0.284971 0.048500 -5.876 4.84e-09 \*\*\*  
## fielding\_e -0.589229 0.037362 -15.771 < 2e-16 \*\*\*  
## fielding\_dp -0.194402 0.022892 -8.492 < 2e-16 \*\*\*  
## batting\_hbp\_bi -0.136648 0.020182 -6.771 1.63e-11 \*\*\*  
## total\_bases 0.388993 0.047997 8.105 8.59e-16 \*\*\*  
## total\_bases\_allowed 0.189220 0.042893 4.411 1.08e-05 \*\*\*  
## HR\_over\_OP -0.101620 0.030242 -3.360 0.000792 \*\*\*  
## walks\_over\_OP 0.161277 0.045189 3.569 0.000366 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7876 on 2249 degrees of freedom  
## Multiple R-squared: 0.3703, Adjusted R-squared: 0.3664   
## F-statistic: 94.48 on 14 and 2249 DF, p-value: < 2.2e-16

paste('MSE equal ', mse(stepwise\_base\_model\_bw\_rm))

## [1] "MSE equal 0.616142519793743"

## Conclusion

It definitely made a difference when the transformation was applied. The only problem that I faced was that my prediction when in the thousands if I used the models created on the transformed data set. After paying a close attention on the Cook's distance for the models' residual, I removed certain observation that led to an improved model.

After testing more than 10 models, using different techniques and transformation, I settled with a model built after I capped outliers, removed variables causing multicollinearity, variables with low p-value, and removed influencial observations.

Here is the model base\_model\_lp\_rm: Target Wins = 32.157432 - 0.035903 \* moneyball\_imp\_test$batting\_2b + 0.068862 \* moneyball\_imp\_test$batting\_3b + 0.044466 \* moneyball\_imp\_test$batting\_bb - 0.016966 \* moneyball\_imp\_test$batting\_so + 0.060647 \* moneyball\_imp\_test$baserun\_sb - 0.050230 \* moneyball\_imp\_test$pitching\_bb - 0.043364 \* moneyball\_imp\_test$fielding\_e - 0.105258 \* moneyball\_imp\_test$fielding\_dp - 4.089404 \* moneyball\_imp\_test$batting\_hbp\_bi + 0.021326 \* moneyball\_imp\_test$total\_bases + 0.011782 \* moneyball\_imp\_test$total\_bases\_allowed + 0.023997 \* moneyball\_imp\_test$walks\_over\_OP + 0.008021 \* moneyball\_imp\_test$SO\_over\_OP

When looking at the Rsquared and Adjusted Rsquared together with the residual plots, the base\_model\_lp\_rm model was not the best model. The stepwise model after removing influencial observation were the best model, but when tested on the test dataset, the numbers were in the thousands. It could be a step I missed, but base\_model\_lp\_rm will be my final model.