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)  
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

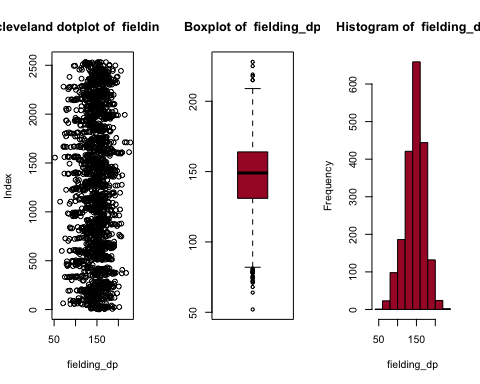
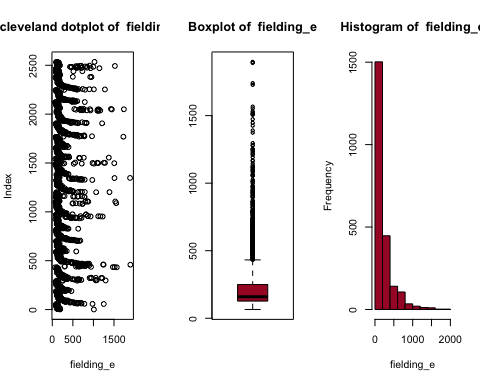
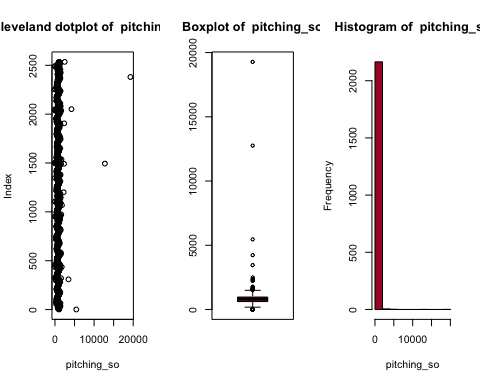
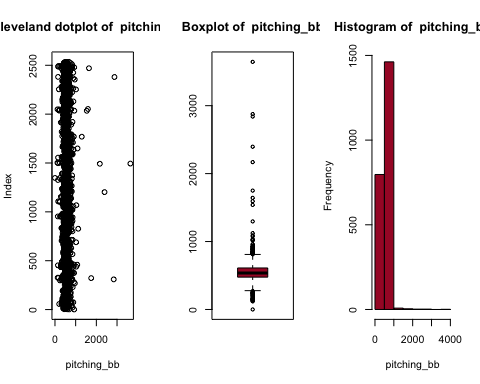
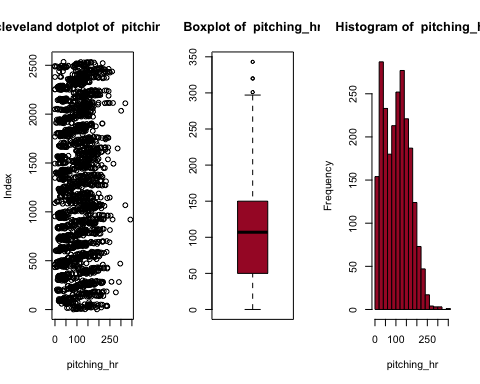
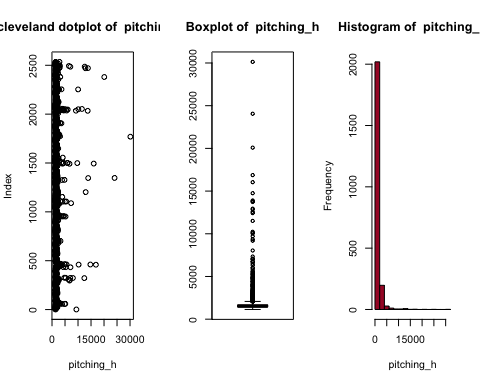
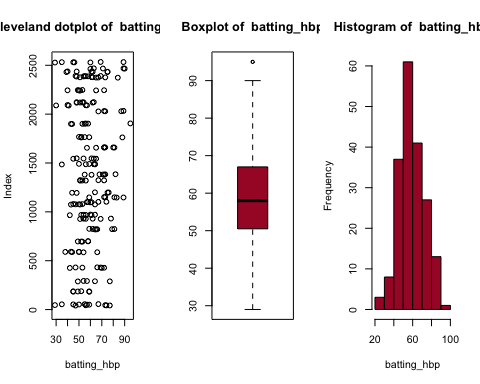
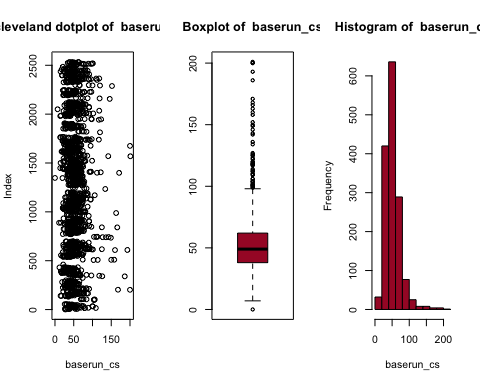
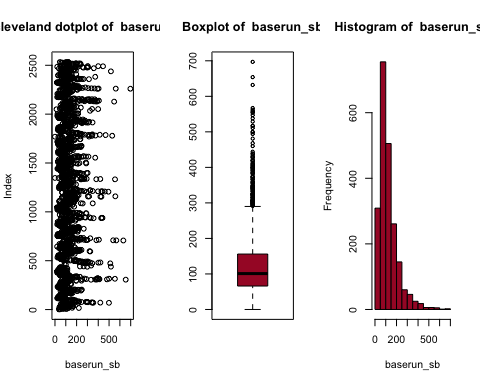
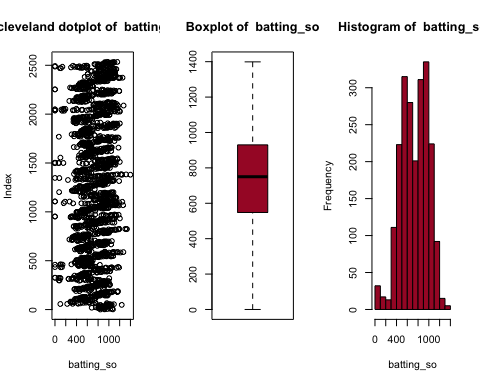
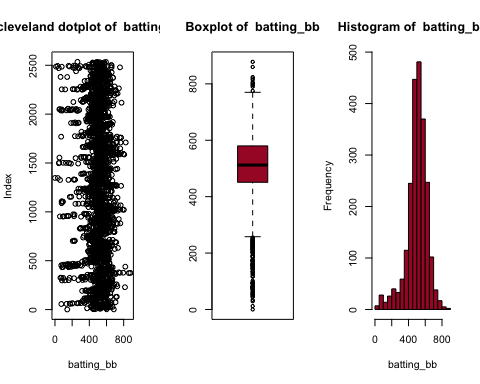
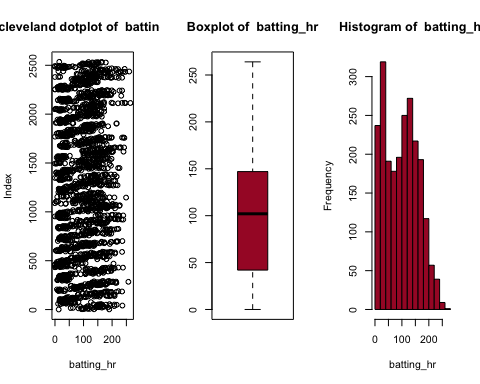
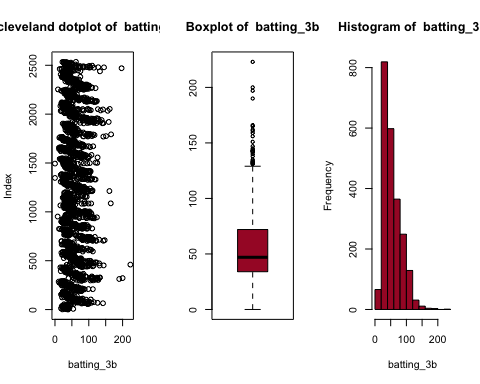
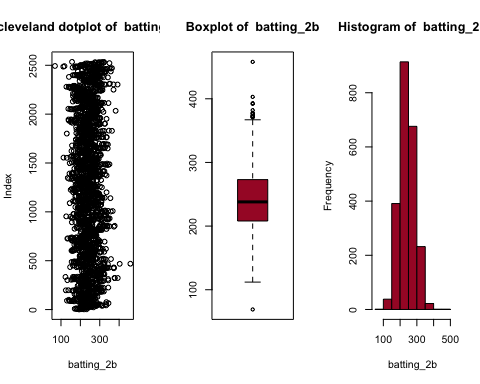
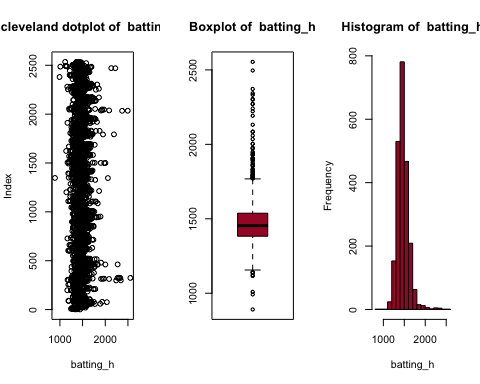
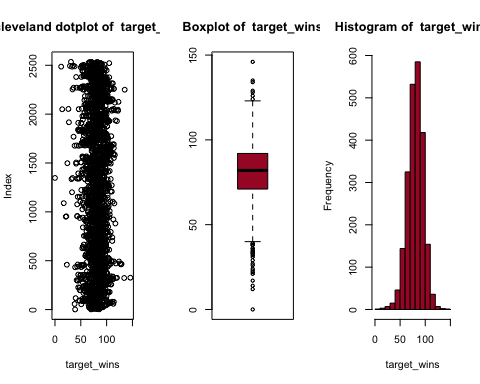
The moneyball dataset has sparked many companies, teams, and organizations to understand and utilize the data they generate/gather. This project highlights many pitfalls that those same individuals fall into simply because they forgot to do the due diligence and prepare the data before modeling.  
This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

## Data Exploration

### Step 1: 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)) {  
   
plot(moneyball[,i], moneyball$index, xlab = colnames(moneyball)[i] , ylab = "Index", main = paste("cleveland dotplot of ",colnames(moneyball)[i]))  
  
boxplot(moneyball[,i], col = "#A71930", main = paste("Boxplot of ",colnames(moneyball)[i]))  
  
hist(  
 moneyball[,i],  
 col = "#A71930",  
 xlab = colnames(moneyball)[i],  
 main = paste("Histogram of ",colnames(moneyball)[i])  
)  
 i = i + 1  
}



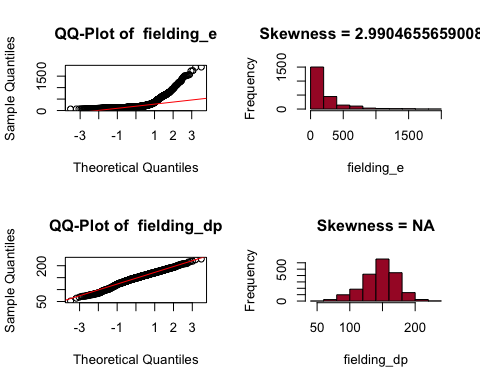
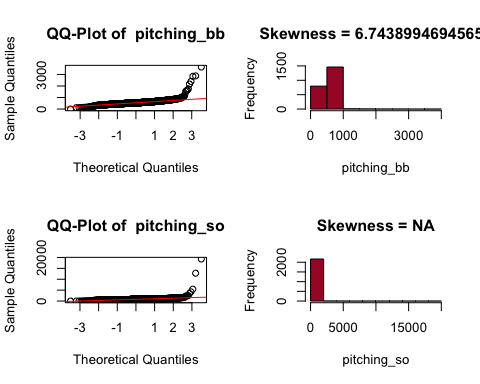
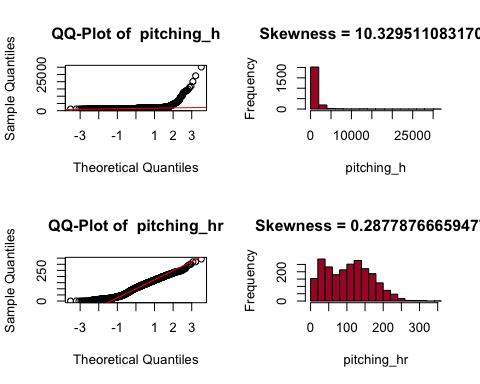
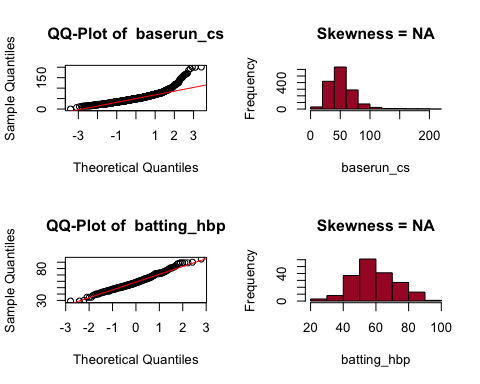
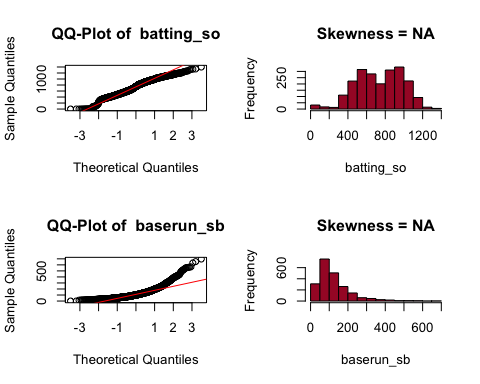
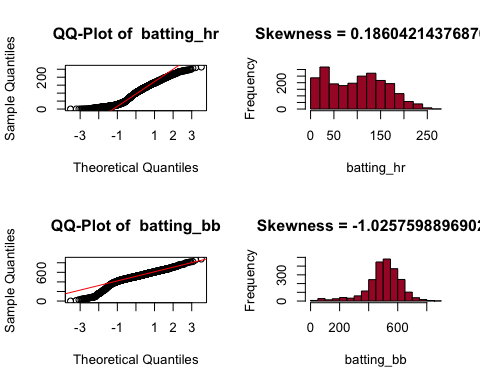
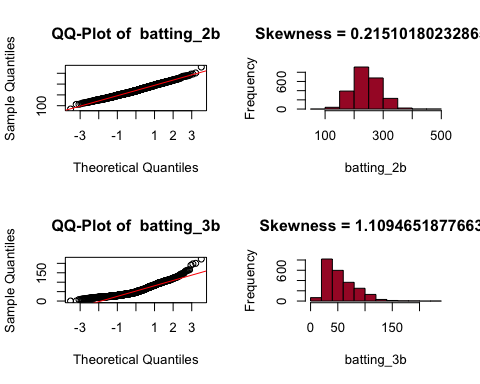
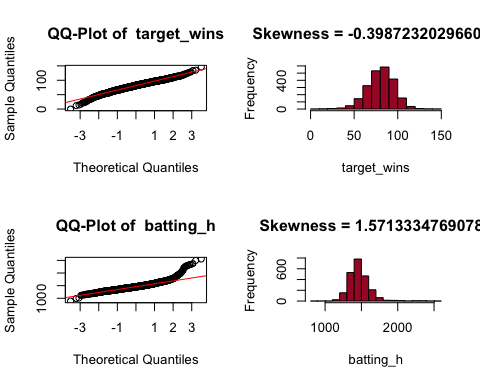
It looks like the outliers are legitmate and we will try Spatial Sign transformation to deal with them.

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

### Step 2: Are the data normally distributed?

From the historgram above we can clearly see that 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(2:17)) {  
 qqnorm(moneyball[,i], main = paste("QQ-Plot of ",colnames(moneyball)[i]));qqline(moneyball[,i], col = 2)  
   
 hist(  
 moneyball[,i],  
 col = "#A71930",  
 xlab = colnames(moneyball)[i],  
 main = paste0("Skewness = ",skewness(moneyball[,i]))  
)  
   
 i = i + 1  
   
}



We would need to try certain transformation to correct for Skewness, with Box-Cox being the number one choice.

### Step 3: 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, I will keep it out until all th eimputation 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 mice package will be the package used to help us with this task. Since we only have numeric values, mice will automatically chose PMM (Predictive Mean Matching) as the method. A great resource to understand this techinique is found [here](https://statisticalhorizons.com/predictive-mean-matching).

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

moneyball\_imp$batting\_hbp <- batting\_hbp  
moneyball\_imp$batting\_hbp\_bi <- batting\_hbp\_bi

### 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\_imp)[findCorrelation(cor(moneyball\_imp))]

## [1] "batting\_hr"

## 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\_imp$batting\_1B <- moneyball\_imp$batting\_h-(moneyball\_imp$batting\_2b + moneyball\_imp$batting\_3b + moneyball\_imp$batting\_hr)  
moneyball\_imp$free\_bases\_num <- if\_else(is.na(moneyball\_imp$batting\_hbp),0,as.numeric(moneyball\_imp$batting\_hbp)) + moneyball\_imp$batting\_bb  
moneyball\_imp$total\_bases <- moneyball\_imp$batting\_1B + 2 \* moneyball\_imp$batting\_2b + 3 \* moneyball\_imp$batting\_3b + 4 \* moneyball\_imp$batting\_hr + moneyball\_imp$batting\_bb + if\_else(is.na(moneyball\_imp$batting\_hbp),0,as.numeric(moneyball\_imp$batting\_hbp)) + moneyball\_imp$baserun\_sb  
moneyball\_imp$total\_bases\_allowed = moneyball\_imp$pitching\_bb + 4 \* moneyball\_imp$pitching\_hr + moneyball\_imp$pitching\_h  
moneyball\_imp$HR\_over\_OP = moneyball\_imp$batting\_hr - moneyball\_imp$pitching\_hr  
moneyball\_imp$walks\_over\_OP = moneyball\_imp$batting\_bb - moneyball\_imp$pitching\_bb  
moneyball\_imp$SO\_over\_OP = moneyball\_imp$pitching\_so - moneyball\_imp$batting\_so  
# make alist of predictors and format them  
  
colnames(moneyball\_imp)

## [1] "index" "target\_wins" "batting\_h"   
## [4] "batting\_2b" "batting\_3b" "batting\_hr"   
## [7] "batting\_bb" "batting\_so" "baserun\_sb"   
## [10] "baserun\_cs" "pitching\_h" "pitching\_hr"   
## [13] "pitching\_bb" "pitching\_so" "fielding\_e"   
## [16] "fielding\_dp" "batting\_hbp" "batting\_hbp\_bi"   
## [19] "batting\_1B" "free\_bases\_num" "total\_bases"   
## [22] "total\_bases\_allowed" "HR\_over\_OP" "walks\_over\_OP"   
## [25] "SO\_over\_OP"

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

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

## 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 0.083191276 -0.036729891 -0.430399494 0.182832186  
## baserun\_sb 0.022217354 0.110667505 0.126436175 -0.199662449  
## baserun\_cs -0.016185369 0.055149220 0.104088789 -0.293656162  
## 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 0.057331050 -0.075891894 -0.236663255 0.077638339  
## fielding\_e -0.009233126 -0.176484759 0.264902478 -0.235150986  
## fielding\_dp 0.018832296 -0.035199171 -0.014962731 0.297600262  
## 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 0.021876208 0.481798618 0.637553115 0.705354607  
## 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 0.021595926 -0.065287353 -0.045652744 -0.005843828  
## batting\_3b batting\_hr batting\_bb batting\_so  
## index -0.005814683 0.051481047 -0.02656724 0.08319128  
## target\_wins 0.142608411 0.176153200 0.23255986 -0.03672989  
## batting\_h 0.427696575 -0.006544685 -0.07246401 -0.43039949  
## batting\_2b -0.107305824 0.435397293 0.25572610 0.18283219  
## batting\_3b 1.000000000 -0.635566946 -0.28723584 -0.67114658  
## batting\_hr -0.635566946 1.000000000 0.51373481 0.72464642  
## batting\_bb -0.287235841 0.513734810 1.00000000 0.38417152  
## batting\_so -0.671146583 0.724646423 0.38417152 1.00000000  
## baserun\_sb 0.537063632 -0.502005173 -0.32211345 -0.29804105  
## baserun\_cs 0.622170881 -0.622742028 -0.34060637 -0.43644472  
## pitching\_h 0.194879411 -0.250145481 -0.44977762 -0.36473064  
## pitching\_hr -0.567836679 0.969371396 0.45955207 0.66702614  
## pitching\_bb -0.002224148 0.136927564 0.48936126 0.05074763  
## pitching\_so -0.263855555 0.194356085 -0.01265517 0.42186287  
## fielding\_e 0.509778447 -0.587339098 -0.65597081 -0.58314626  
## fielding\_dp -0.457575674 0.517351743 0.38525011 0.32051339  
## batting\_hbp\_bi -0.265544426 0.392199209 0.10305838 0.39665688  
## batting\_1B 0.600399234 -0.497294855 -0.35312165 -0.74394613  
## free\_bases\_num -0.316009005 0.553966941 0.99101046 0.42825264  
## total\_bases 0.036084290 0.596628779 0.55152091 0.20788926  
## total\_bases\_allowed 0.092039617 -0.062551344 -0.30004852 -0.23332402  
## HR\_over\_OP -0.243354524 0.074559388 0.19441460 0.19786816  
## walks\_over\_OP -0.231156161 0.266798215 0.27356493 0.25647557  
## SO\_over\_OP 0.044755858 -0.148115027 -0.20604861 -0.03502684  
## baserun\_sb baserun\_cs pitching\_h pitching\_hr  
## index 0.02221735 -0.01618537 0.01710315 0.05098590  
## target\_wins 0.11066751 0.05514922 -0.10993705 0.18901373  
## batting\_h 0.12643617 0.10408879 0.30269371 0.07285312  
## batting\_2b -0.19966245 -0.29365616 0.02369219 0.45455082  
## batting\_3b 0.53706363 0.62217088 0.19487941 -0.56783668  
## batting\_hr -0.50200517 -0.62274203 -0.25014548 0.96937140  
## batting\_bb -0.32211345 -0.34060637 -0.44977762 0.45955207  
## batting\_so -0.29804105 -0.43644472 -0.36473064 0.66702614  
## baserun\_sb 1.00000000 0.81438594 0.15531307 -0.45370683  
## baserun\_cs 0.81438594 1.00000000 0.13729046 -0.58171865  
## pitching\_h 0.15531307 0.13729046 1.00000000 -0.14161276  
## pitching\_hr -0.45370683 -0.58171865 -0.14161276 1.00000000  
## pitching\_bb 0.02004583 -0.01295159 0.32067616 0.22193750  
## pitching\_so 0.02345640 -0.05140041 0.26875318 0.21454801  
## fielding\_e 0.57615999 0.54419921 0.66775901 -0.49314447  
## fielding\_dp -0.63152895 -0.63039816 -0.08103926 0.48713946  
## batting\_hbp\_bi -0.13489925 -0.21078191 -0.06445004 0.35794984  
## batting\_1B 0.33365307 0.38096961 0.40612014 -0.41549520  
## free\_bases\_num -0.33263968 -0.36063911 -0.44800796 0.49652206  
## total\_bases 0.01876159 -0.13232666 -0.09806792 0.62309982  
## total\_bases\_allowed 0.07564125 0.03307542 0.97499650 0.05669475  
## HR\_over\_OP -0.17110614 -0.13555426 -0.42822141 -0.17264012  
## walks\_over\_OP -0.28408463 -0.26273075 -0.71949139 0.12897043  
## SO\_over\_OP 0.17488626 0.16158030 0.47861739 -0.09704586  
## pitching\_bb pitching\_so fielding\_e fielding\_dp  
## index -0.015287513 0.057331050 -0.009233126 0.01883230  
## target\_wins 0.124174536 -0.075891894 -0.176484759 -0.03519917  
## batting\_h 0.094193027 -0.236663255 0.264902478 -0.01496273  
## batting\_2b 0.178054204 0.077638339 -0.235150986 0.29760026  
## batting\_3b -0.002224148 -0.263855555 0.509778447 -0.45757567  
## batting\_hr 0.136927564 0.194356085 -0.587339098 0.51735174  
## batting\_bb 0.489361263 -0.012655173 -0.655970815 0.38525011  
## batting\_so 0.050747626 0.421862874 -0.583146261 0.32051339  
## baserun\_sb 0.020045833 0.023456395 0.576159995 -0.63152895  
## baserun\_cs -0.012951588 -0.051400409 0.544199205 -0.63039816  
## pitching\_h 0.320676162 0.268753181 0.667759010 -0.08103926  
## pitching\_hr 0.221937505 0.214548013 -0.493144466 0.48713946  
## pitching\_bb 1.000000000 0.487517295 -0.022837561 0.15216542  
## pitching\_so 0.487517295 1.000000000 -0.027032833 0.13197958  
## fielding\_e -0.022837561 -0.027032833 1.000000000 -0.52099084  
## fielding\_dp 0.152165415 0.131979581 -0.520990838 1.00000000  
## batting\_hbp\_bi -0.016906833 0.134856127 -0.185315470 0.11414223  
## batting\_1B -0.022820326 -0.327702412 0.547816415 -0.26861907  
## free\_bases\_num 0.476195183 0.005684899 -0.665319984 0.39108911  
## total\_bases 0.349856050 -0.022484608 -0.248053870 0.21359723  
## total\_bases\_allowed 0.459579945 0.350064125 0.557252830 0.02103147  
## HR\_over\_OP -0.351988418 -0.091665561 -0.353210656 0.09690633  
## walks\_over\_OP -0.704942270 -0.547994982 -0.508313405 0.14549501  
## SO\_over\_OP 0.512001527 0.891326770 0.261795901 -0.01479035  
## batting\_hbp\_bi batting\_1B free\_bases\_num total\_bases  
## index 0.047332196 -0.04707442 -0.019063695 0.02187621  
## target\_wins 0.002610647 0.21743014 0.228098279 0.48179862  
## batting\_h 0.019594018 0.82758476 -0.068377971 0.63755312  
## batting\_2b 0.361922796 0.08700989 0.297591911 0.70535461  
## batting\_3b -0.265544426 0.60039923 -0.316009005 0.03608429  
## batting\_hr 0.392199209 -0.49729485 0.553966941 0.59662878  
## batting\_bb 0.103058382 -0.35312165 0.991010459 0.55152091  
## batting\_so 0.396656876 -0.74394613 0.428252642 0.20788926  
## baserun\_sb -0.134899249 0.33365307 -0.332639681 0.01876159  
## baserun\_cs -0.210781913 0.38096961 -0.360639111 -0.13232666  
## pitching\_h -0.064450039 0.40612014 -0.448007961 -0.09806792  
## pitching\_hr 0.357949841 -0.41549520 0.496522065 0.62309982  
## pitching\_bb -0.016906833 -0.02282033 0.476195183 0.34985605  
## pitching\_so 0.134856127 -0.32770241 0.005684899 -0.02248461  
## fielding\_e -0.185315470 0.54781641 -0.665319984 -0.24805387  
## fielding\_dp 0.114142229 -0.26861907 0.391089112 0.21359723  
## batting\_hbp\_bi 1.000000000 -0.23605172 0.231848863 0.29611106  
## batting\_1B -0.236051718 1.00000000 -0.376395883 0.17095897  
## free\_bases\_num 0.231848863 -0.37639588 1.000000000 0.57838032  
## total\_bases 0.296111061 0.17095897 0.578380323 1.00000000  
## total\_bases\_allowed -0.003909755 0.31851323 -0.293643548 0.04994171  
## HR\_over\_OP 0.119531251 -0.30736736 0.205656303 -0.13719528  
## walks\_over\_OP 0.102464739 -0.26202481 0.280775130 0.06268227  
## SO\_over\_OP -0.049694011 0.01078151 -0.207874940 -0.12873605  
## 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 -0.233324016 0.19786816 0.256475569  
## baserun\_sb 0.075641250 -0.17110614 -0.284084626  
## baserun\_cs 0.033075422 -0.13555426 -0.262730748  
## 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 0.350064125 -0.09166556 -0.547994982  
## fielding\_e 0.557252830 -0.35321066 -0.508313405  
## fielding\_dp 0.021031466 0.09690633 0.145495010  
## 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 0.049941712 -0.13719528 0.062682267  
## 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 0.502536346 -0.19998144 -0.732286753  
## SO\_over\_OP  
## index 0.021595926  
## target\_wins -0.065287353  
## batting\_h -0.045652744  
## batting\_2b -0.005843828  
## batting\_3b 0.044755858  
## batting\_hr -0.148115027  
## batting\_bb -0.206048609  
## batting\_so -0.035026843  
## baserun\_sb 0.174886262  
## baserun\_cs 0.161580298  
## pitching\_h 0.478617387  
## pitching\_hr -0.097045860  
## pitching\_bb 0.512001527  
## pitching\_so 0.891326770  
## fielding\_e 0.261795901  
## fielding\_dp -0.014790353  
## batting\_hbp\_bi -0.049694011  
## batting\_1B 0.010781508  
## free\_bases\_num -0.207874940  
## total\_bases -0.128736052  
## total\_bases\_allowed 0.502536346  
## HR\_over\_OP -0.199981441  
## walks\_over\_OP -0.732286753  
## SO\_over\_OP 1.000000000

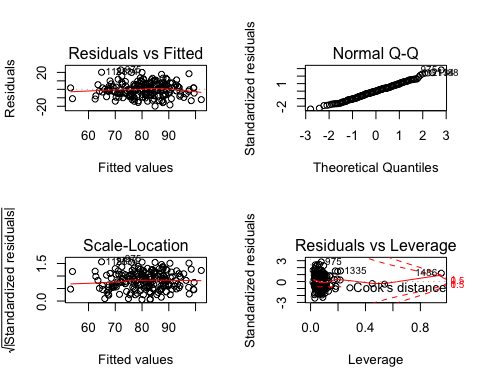
## Build a Model

Let's test a model to establish a baseline

str(moneyball\_imp)

## 'data.frame': 2276 obs. of 24 variables:  
## $ index : int 1 2 3 4 5 6 7 8 11 12 ...  
## $ target\_wins : int 39 70 86 70 82 75 80 85 86 76 ...  
## $ batting\_h : int 1445 1339 1377 1387 1297 1279 1244 1273 1391 1271 ...  
## $ batting\_2b : int 194 219 232 209 186 200 179 171 197 213 ...  
## $ batting\_3b : int 39 22 35 38 27 36 54 37 40 18 ...  
## $ batting\_hr : int 13 190 137 96 102 92 122 115 114 96 ...  
## $ batting\_bb : int 143 685 602 451 472 443 525 456 447 441 ...  
## $ batting\_so : int 842 1075 917 922 920 973 1062 1027 922 827 ...  
## $ baserun\_sb : int 341 37 46 43 49 107 80 40 69 72 ...  
## $ baserun\_cs : int 193 28 27 30 39 59 54 36 27 34 ...  
## $ pitching\_h : int 9364 1347 1377 1396 1297 1279 1244 1281 1391 1271 ...  
## $ pitching\_hr : int 84 191 137 97 102 92 122 116 114 96 ...  
## $ pitching\_bb : int 927 689 602 454 472 443 525 459 447 441 ...  
## $ pitching\_so : int 5456 1082 917 928 920 973 1062 1033 922 827 ...  
## $ fielding\_e : int 1011 193 175 164 138 123 136 112 127 131 ...  
## $ fielding\_dp : int 162 155 153 156 168 149 186 136 169 159 ...  
## $ batting\_hbp\_bi : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ batting\_1B : int 1199 908 973 1044 982 951 889 950 1040 944 ...  
## $ free\_bases\_num : num 143 685 602 451 472 443 525 456 447 441 ...  
## $ total\_bases : num 2240 2894 2738 2454 2364 ...  
## $ total\_bases\_allowed: num 10627 2800 2527 2238 2177 ...  
## $ HR\_over\_OP : int -71 -1 0 -1 0 0 0 -1 0 0 ...  
## $ walks\_over\_OP : int -784 -4 0 -3 0 0 0 -3 0 0 ...  
## $ SO\_over\_OP : int 4614 7 0 6 0 0 0 6 0 0 ...

base\_model\_all <- lm(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, data = moneyball\_imp)  
par(mfrow=c(2,2))  
plot(base\_model\_all)



summary(base\_model\_all)

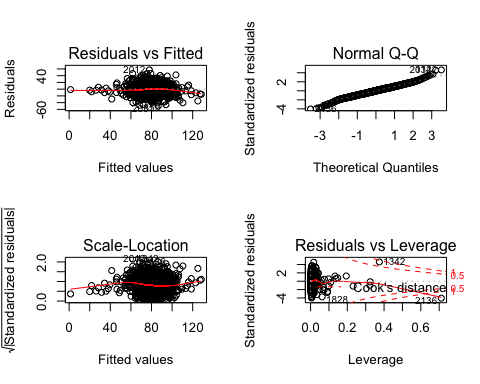
##   
## Call:  
## lm(formula = target\_wins ~ batting\_h + batting\_2b + batting\_3b +   
## batting\_hr + batting\_bb + batting\_so + baserun\_sb + baserun\_cs +   
## pitching\_h + pitching\_hr + pitching\_bb + pitching\_so + fielding\_e +   
## fielding\_dp + batting\_hbp + batting\_hbp\_bi + batting\_1B +   
## free\_bases\_num + total\_bases + total\_bases\_allowed + HR\_over\_OP +   
## walks\_over\_OP + SO\_over\_OP, data = moneyball\_imp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.8708 -5.6564 -0.0599 5.2545 22.9274   
##   
## Coefficients: (8 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 60.28826 19.67842 3.064 0.00253 \*\*   
## batting\_h 1.91348 2.76139 0.693 0.48927   
## batting\_2b 0.02639 0.03029 0.871 0.38484   
## batting\_3b -0.10118 0.07751 -1.305 0.19348   
## batting\_hr -4.84371 10.50851 -0.461 0.64542   
## batting\_bb -4.45969 3.63624 -1.226 0.22167   
## batting\_so 0.34196 2.59876 0.132 0.89546   
## baserun\_sb 0.03304 0.02867 1.152 0.25071   
## baserun\_cs -0.01104 0.07143 -0.155 0.87730   
## pitching\_h -1.89096 2.76095 -0.685 0.49432   
## pitching\_hr 4.93043 10.50664 0.469 0.63946   
## pitching\_bb 4.51089 3.63372 1.241 0.21612   
## pitching\_so -0.37364 2.59705 -0.144 0.88577   
## fielding\_e -0.17204 0.04140 -4.155 5.08e-05 \*\*\*  
## fielding\_dp -0.10819 0.03654 -2.961 0.00349 \*\*   
## batting\_hbp 0.08247 0.04960 1.663 0.09815 .   
## batting\_hbp\_bi NA NA NA NA   
## batting\_1B NA NA NA NA   
## free\_bases\_num NA NA NA NA   
## total\_bases NA NA NA NA   
## total\_bases\_allowed NA NA NA NA   
## 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: 8.467 on 175 degrees of freedom  
## (2085 observations deleted due to missingness)  
## Multiple R-squared: 0.5501, Adjusted R-squared: 0.5116   
## F-statistic: 14.27 on 15 and 175 DF, p-value: < 2.2e-16

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

## [1] "MSE equal 65.6852879651226"

Though R-squared and adjusted R-square is high, we can clearly see that this model dropping observations. Let's try to forget about the new additions, and build a model without them.

moneyball\_orig <- moneyball\_imp[,1:17]  
base\_model\_orig <-  
 lm(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, data = moneyball\_orig)  
 par(mfrow = c(2, 2))  
 plot(base\_model\_orig)



summary(base\_model\_orig)

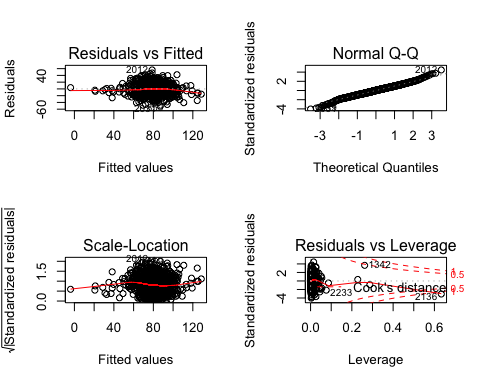
##   
## Call:  
## lm(formula = target\_wins ~ batting\_h + batting\_2b + batting\_3b +   
## batting\_hr + batting\_bb + batting\_so + baserun\_sb + baserun\_cs +   
## pitching\_h + pitching\_hr + pitching\_bb + pitching\_so + fielding\_e +   
## fielding\_dp, data = moneyball\_orig)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -50.437 -8.273 0.109 8.115 57.063   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34.5833869 5.2232323 6.621 4.44e-11 \*\*\*  
## batting\_h 0.0434011 0.0035801 12.123 < 2e-16 \*\*\*  
## batting\_2b -0.0203630 0.0089278 -2.281 0.02265 \*   
## batting\_3b 0.0295276 0.0166056 1.778 0.07551 .   
## batting\_hr 0.0604145 0.0265592 2.275 0.02302 \*   
## batting\_bb 0.0140708 0.0056443 2.493 0.01274 \*   
## batting\_so -0.0168623 0.0025071 -6.726 2.20e-11 \*\*\*  
## baserun\_sb 0.0529984 0.0052813 10.035 < 2e-16 \*\*\*  
## baserun\_cs -0.0047414 0.0104140 -0.455 0.64894   
## pitching\_h 0.0011718 0.0003812 3.074 0.00214 \*\*   
## pitching\_hr 0.0198220 0.0235832 0.841 0.40071   
## pitching\_bb -0.0055801 0.0040211 -1.388 0.16536   
## pitching\_so 0.0026248 0.0008980 2.923 0.00350 \*\*   
## fielding\_e -0.0407587 0.0026676 -15.279 < 2e-16 \*\*\*  
## fielding\_dp -0.1067389 0.0130221 -8.197 4.09e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.67 on 2261 degrees of freedom  
## Multiple R-squared: 0.3573, Adjusted R-squared: 0.3533   
## F-statistic: 89.77 on 14 and 2261 DF, p-value: < 2.2e-16

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

## [1] "MSE equal 159.414751654005"

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.  
Let's build another model including lon those with low p-Values.

base\_model\_lp <-  
 lm(target\_wins ~ batting\_h + batting\_2b + batting\_hr + batting\_bb + batting\_so + baserun\_sb + pitching\_h + pitching\_so + fielding\_e + fielding\_dp, data = moneyball\_orig)  
 par(mfrow = c(2, 2))  
 plot(base\_model\_lp)



summary(base\_model\_lp)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_h + batting\_2b + batting\_hr +   
## batting\_bb + batting\_so + baserun\_sb + pitching\_h + pitching\_so +   
## fielding\_e + fielding\_dp, data = moneyball\_orig)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -51.044 -8.404 0.170 8.266 56.224   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 33.7481921 5.0952471 6.623 4.37e-11 \*\*\*  
## batting\_h 0.0458675 0.0033260 13.790 < 2e-16 \*\*\*  
## batting\_2b -0.0215281 0.0088402 -2.435 0.01496 \*   
## batting\_hr 0.0771546 0.0089649 8.606 < 2e-16 \*\*\*  
## batting\_bb 0.0080930 0.0030417 2.661 0.00785 \*\*   
## batting\_so -0.0165511 0.0023941 -6.913 6.13e-12 \*\*\*  
## baserun\_sb 0.0527059 0.0041691 12.642 < 2e-16 \*\*\*  
## pitching\_h 0.0008875 0.0003305 2.686 0.00729 \*\*   
## pitching\_so 0.0018158 0.0006642 2.734 0.00631 \*\*   
## fielding\_e -0.0405882 0.0026636 -15.238 < 2e-16 \*\*\*  
## fielding\_dp -0.1064669 0.0128053 -8.314 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.67 on 2265 degrees of freedom  
## Multiple R-squared: 0.3557, Adjusted R-squared: 0.3529   
## F-statistic: 125.1 on 10 and 2265 DF, p-value: < 2.2e-16

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

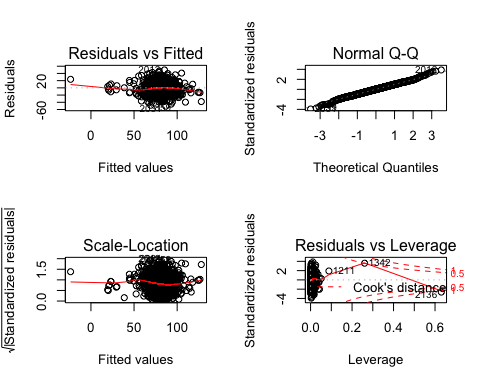
## [1] "MSE equal 159.788274102743"

Lets remove variables causing multicollinearity using findCorrelation().

to\_rm <- colnames(cor(moneyball\_imp)[,findCorrelation(cor(moneyball\_imp))])  
to\_rm

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

base\_model\_noCol <-  
 lm(target\_wins ~ batting\_h + batting\_2b + batting\_bb + batting\_so + baserun\_sb + pitching\_so + fielding\_e + fielding\_dp, data = moneyball\_orig)  
 par(mfrow = c(2, 2))  
 plot(base\_model\_noCol)



summary(base\_model\_noCol)

##   
## Call:  
## lm(formula = target\_wins ~ batting\_h + batting\_2b + batting\_bb +   
## batting\_so + baserun\_sb + pitching\_so + fielding\_e + fielding\_dp,   
## data = moneyball\_orig)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -50.696 -8.530 0.266 8.443 49.730   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.9492583 4.3944541 2.036 0.041817 \*   
## batting\_h 0.0574802 0.0031309 18.359 < 2e-16 \*\*\*  
## batting\_2b -0.0192319 0.0089927 -2.139 0.032573 \*   
## batting\_bb 0.0145786 0.0029924 4.872 1.18e-06 \*\*\*  
## batting\_so -0.0027687 0.0017242 -1.606 0.108471   
## baserun\_sb 0.0384923 0.0038631 9.964 < 2e-16 \*\*\*  
## pitching\_so 0.0021644 0.0005971 3.624 0.000296 \*\*\*  
## fielding\_e -0.0346238 0.0021304 -16.253 < 2e-16 \*\*\*  
## fielding\_dp -0.0829051 0.0126539 -6.552 7.01e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.9 on 2267 degrees of freedom  
## Multiple R-squared: 0.3318, Adjusted R-squared: 0.3294   
## F-statistic: 140.7 on 8 and 2267 DF, p-value: < 2.2e-16

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

## [1] "MSE equal 165.726715773464"

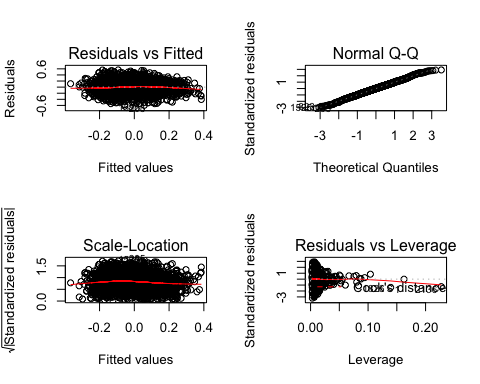
Though the rsquared value went down, there are some improvements on the Cook's distance chart. Now let's try to use use the caret package to apply the transformations we discussed earlier in our exploration phase.

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(moneyball\_imp, method = c("center","scale","spatialSign","BoxCox"))  
transformed <- predict(trans, moneyball\_imp)  
head(transformed)

## index target\_wins batting\_h batting\_2b batting\_3b batting\_hr  
## 1 -0.1318993 -0.16011149 -0.004379949 -0.06096088 -0.03510189 -0.086331170  
## 2 -0.4250602 -0.13382356 -0.187695179 -0.08665517 -0.23248989 0.291632061  
## 3 -0.6210943 0.09475494 -0.178878540 -0.04563309 -0.20768106 0.176935877  
## 4 -0.6188686 -0.19625924 -0.154489134 -0.19106820 -0.17688827 -0.017091280  
## 5 -0.5063962 0.01805238 -0.319669581 -0.28162339 -0.23780046 0.009275424  
## 6 -0.4935417 -0.08451807 -0.353883651 -0.20038551 -0.15840652 -0.028903842  
## batting\_bb batting\_so baserun\_sb baserun\_cs pitching\_h pitching\_hr  
## 1 -0.17640063 0.02781425 0.12432740 0.13906960 0.1905490 -0.02136294  
## 2 0.29212698 0.27499076 -0.19255528 -0.18175300 -0.2150879 0.27184477  
## 3 0.23460971 0.21938185 -0.25659560 -0.27220563 -0.2617840 0.14631474  
## 4 -0.11807825 0.22517139 -0.26517013 -0.25531711 -0.2295180 -0.04065476  
## 5 -0.05666889 0.18292936 -0.20353764 -0.16809172 -0.3393062 -0.01419003  
## 6 -0.10974825 0.22833966 -0.06537484 -0.07417815 -0.3622594 -0.05137723  
## pitching\_bb pitching\_so fielding\_e fielding\_dp batting\_hbp\_bi  
## 1 0.13567551 0.51684030 0.11668974 0.04151368 -0.01826204  
## 2 0.15969379 0.09789088 0.05223896 0.08465340 -0.05911322  
## 3 0.08438394 0.05643188 0.01986019 0.10362578 -0.08670498  
## 4 -0.17050683 0.06223305 -0.02062926 0.13446734 -0.08669262  
## 5 -0.11452060 0.04761825 -0.11535572 0.21412570 -0.07116500  
## 6 -0.15202961 0.06901141 -0.18573160 0.05050631 -0.06956880  
## batting\_1B free\_bases\_num total\_bases total\_bases\_allowed HR\_over\_OP  
## 1 0.06813822 -0.1747816 -0.10155914 0.1795229556 -0.25949334  
## 2 -0.33854592 0.2777281 0.08330791 0.1104436086 0.06581890  
## 3 -0.24824669 0.2179013 -0.02197071 0.0006254609 0.11552019  
## 4 -0.02823218 -0.1267600 -0.28438173 -0.2358438843 0.09652686  
## 5 -0.17867725 -0.0647117 -0.30171038 -0.2443840403 0.09481570  
## 6 -0.26206446 -0.1163740 -0.28530389 -0.3170975749 0.09268903  
## walks\_over\_OP SO\_over\_OP  
## 1 -0.29310885 0.55579097  
## 2 0.06145451 -0.02960250  
## 3 0.09773791 -0.04749524  
## 4 0.09202568 -0.04399574  
## 5 0.08022051 -0.03898275  
## 6 0.07842121 -0.03810839

trans\_model\_all <-  
 lm(target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb + batting\_so + baserun\_sb + baserun\_cs + pitching\_h + pitching\_hr + pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B + free\_bases\_num + 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\_h +   
## pitching\_hr + pitching\_bb + pitching\_so + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + batting\_1B + free\_bases\_num + total\_bases +   
## total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP + SO\_over\_OP,   
## data = transformed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.62430 -0.12992 0.00302 0.12901 0.56473   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.002023 0.004503 0.449 0.653347   
## batting\_h -0.157125 0.218126 -0.720 0.471391   
## batting\_2b -0.154427 0.092172 -1.675 0.093991 .   
## batting\_3b 0.151461 0.082784 1.830 0.067444 .   
## batting\_bb -0.465134 0.724456 -0.642 0.520909   
## batting\_so -0.550649 0.094236 -5.843 5.86e-09 \*\*\*  
## baserun\_sb 0.062377 0.106454 0.586 0.557968   
## baserun\_cs -0.002407 0.037904 -0.064 0.949370   
## pitching\_h -0.122984 0.118058 -1.042 0.297650   
## pitching\_hr -0.105658 0.208575 -0.507 0.612505   
## pitching\_bb -0.367891 0.127904 -2.876 0.004061 \*\*   
## pitching\_so 0.583123 0.175578 3.321 0.000911 \*\*\*  
## fielding\_e -0.526787 0.035344 -14.905 < 2e-16 \*\*\*  
## fielding\_dp -0.174484 0.021579 -8.086 9.96e-16 \*\*\*  
## batting\_hbp\_bi -0.219031 0.098283 -2.229 0.025940 \*   
## batting\_1B 0.133992 0.158747 0.844 0.398725   
## free\_bases\_num 0.660590 0.745334 0.886 0.375550   
## total\_bases 0.729920 0.309296 2.360 0.018363 \*   
## total\_bases\_allowed 0.066528 0.087480 0.760 0.447039   
## HR\_over\_OP -0.051305 0.072371 -0.709 0.478450   
## 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: 0.1948 on 2256 degrees of freedom  
## Multiple R-squared: 0.3021, Adjusted R-squared: 0.2963   
## F-statistic: 51.41 on 19 and 2256 DF, p-value: < 2.2e-16

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

## [1] "MSE equal 0.0376220132488467"

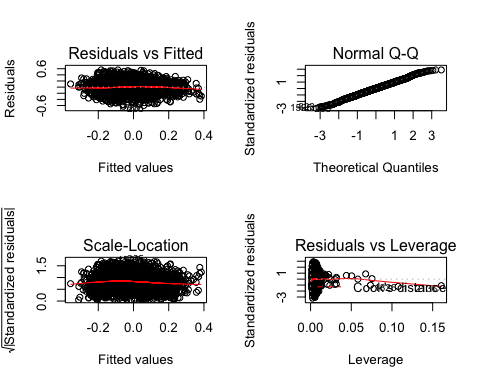
Looking at Cook's Distance, it's clear that we have influential data, but the other charts look right where they should be.

Let's try, stepwise approach. 1. Both direction

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

## Start: AIC=-7425.66  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_h + pitching\_hr +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + free\_bases\_num + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP + walks\_over\_OP + SO\_over\_OP  
##   
##   
## Step: AIC=-7425.66  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_h + pitching\_hr +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + free\_bases\_num + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP + walks\_over\_OP  
##   
##   
## Step: AIC=-7425.66  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_h + pitching\_hr +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + free\_bases\_num + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - baserun\_cs 1 0.0002 85.628 -7427.7  
## - pitching\_hr 1 0.0097 85.637 -7427.4  
## - baserun\_sb 1 0.0130 85.641 -7427.3  
## - batting\_bb 1 0.0156 85.643 -7427.2  
## - HR\_over\_OP 1 0.0191 85.647 -7427.2  
## - batting\_h 1 0.0197 85.647 -7427.1  
## - total\_bases\_allowed 1 0.0220 85.650 -7427.1  
## - batting\_1B 1 0.0270 85.655 -7426.9  
## - free\_bases\_num 1 0.0298 85.658 -7426.9  
## - pitching\_h 1 0.0412 85.669 -7426.6  
## <none> 85.628 -7425.7  
## - batting\_2b 1 0.1065 85.734 -7424.8  
## - batting\_3b 1 0.1271 85.755 -7424.3  
## - batting\_hbp\_bi 1 0.1885 85.816 -7422.7  
## - total\_bases 1 0.2114 85.839 -7422.0  
## - pitching\_bb 1 0.3140 85.942 -7419.3  
## - pitching\_so 1 0.4187 86.046 -7416.6  
## - batting\_so 1 1.2960 86.924 -7393.5  
## - fielding\_dp 1 2.4816 88.109 -7362.6  
## - fielding\_e 1 8.4317 94.059 -7213.9  
##   
## Step: AIC=-7427.65  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + pitching\_h + pitching\_hr + pitching\_bb +   
## pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + free\_bases\_num + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - pitching\_hr 1 0.0096 85.638 -7429.4  
## - baserun\_sb 1 0.0130 85.641 -7429.3  
## - batting\_bb 1 0.0156 85.643 -7429.2  
## - HR\_over\_OP 1 0.0191 85.647 -7429.1  
## - batting\_h 1 0.0199 85.648 -7429.1  
## - total\_bases\_allowed 1 0.0231 85.651 -7429.0  
## - batting\_1B 1 0.0277 85.656 -7428.9  
## - free\_bases\_num 1 0.0298 85.658 -7428.9  
## - pitching\_h 1 0.0423 85.670 -7428.5  
## <none> 85.628 -7427.7  
## - batting\_2b 1 0.1071 85.735 -7426.8  
## - batting\_3b 1 0.1274 85.755 -7426.3  
## + baserun\_cs 1 0.0002 85.628 -7425.7  
## - batting\_hbp\_bi 1 0.1885 85.816 -7424.6  
## - total\_bases 1 0.2113 85.839 -7424.0  
## - pitching\_bb 1 0.3149 85.943 -7421.3  
## - pitching\_so 1 0.4185 86.046 -7418.6  
## - batting\_so 1 1.2974 86.925 -7395.4  
## - fielding\_dp 1 2.4847 88.113 -7364.6  
## - fielding\_e 1 8.4734 94.101 -7214.9  
##   
## Step: AIC=-7429.4  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + pitching\_h + pitching\_bb + pitching\_so +   
## fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B +   
## free\_bases\_num + total\_bases + total\_bases\_allowed + HR\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - HR\_over\_OP 1 0.0095 85.647 -7431.1  
## - batting\_h 1 0.0157 85.653 -7431.0  
## - batting\_bb 1 0.0173 85.655 -7430.9  
## - total\_bases\_allowed 1 0.0284 85.666 -7430.6  
## - free\_bases\_num 1 0.0368 85.674 -7430.4  
## - batting\_1B 1 0.0575 85.695 -7429.9  
## - pitching\_h 1 0.0575 85.695 -7429.9  
## <none> 85.638 -7429.4  
## + pitching\_hr 1 0.0096 85.628 -7427.7  
## + baserun\_cs 1 0.0001 85.637 -7427.4  
## - baserun\_sb 1 0.1819 85.819 -7426.6  
## - batting\_2b 1 0.1856 85.823 -7426.5  
## - batting\_hbp\_bi 1 0.1859 85.823 -7426.5  
## - pitching\_bb 1 0.3081 85.946 -7423.2  
## - pitching\_so 1 0.4138 86.051 -7420.4  
## - total\_bases 1 0.7315 86.369 -7412.0  
## - batting\_3b 1 0.8476 86.485 -7409.0  
## - batting\_so 1 1.3026 86.940 -7397.0  
## - fielding\_dp 1 2.5208 88.158 -7365.4  
## - fielding\_e 1 8.6348 94.272 -7212.8  
##   
## Step: AIC=-7431.14  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + pitching\_h + pitching\_bb + pitching\_so +   
## fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B +   
## free\_bases\_num + total\_bases + total\_bases\_allowed  
##   
## Df Sum of Sq RSS AIC  
## - batting\_h 1 0.0182 85.665 -7432.7  
## - batting\_bb 1 0.0183 85.665 -7432.7  
## - total\_bases\_allowed 1 0.0234 85.670 -7432.5  
## - free\_bases\_num 1 0.0361 85.683 -7432.2  
## - pitching\_h 1 0.0485 85.695 -7431.9  
## - batting\_1B 1 0.0561 85.703 -7431.7  
## <none> 85.647 -7431.1  
## + HR\_over\_OP 1 0.0095 85.638 -7429.4  
## + baserun\_cs 1 0.0002 85.647 -7429.1  
## + pitching\_hr 1 0.0001 85.647 -7429.1  
## - baserun\_sb 1 0.1724 85.819 -7428.6  
## - batting\_hbp\_bi 1 0.1893 85.836 -7428.1  
## - batting\_2b 1 0.1951 85.842 -7428.0  
## - pitching\_bb 1 0.3182 85.965 -7424.7  
## - pitching\_so 1 0.4212 86.068 -7422.0  
## - total\_bases 1 0.7882 86.435 -7412.3  
## - batting\_3b 1 0.8382 86.485 -7411.0  
## - batting\_so 1 1.3090 86.956 -7398.6  
## - fielding\_dp 1 2.5113 88.158 -7367.4  
## - fielding\_e 1 8.8032 94.450 -7210.5  
##   
## Step: AIC=-7432.66  
## target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + pitching\_h + pitching\_bb + pitching\_so + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + batting\_1B + free\_bases\_num +   
## total\_bases + total\_bases\_allowed  
##   
## Df Sum of Sq RSS AIC  
## - total\_bases\_allowed 1 0.0156 85.681 -7434.2  
## - batting\_bb 1 0.0177 85.683 -7434.2  
## - free\_bases\_num 1 0.0382 85.703 -7433.6  
## - pitching\_h 1 0.0479 85.713 -7433.4  
## <none> 85.665 -7432.7  
## - batting\_1B 1 0.0930 85.758 -7432.2  
## + batting\_h 1 0.0182 85.647 -7431.1  
## + HR\_over\_OP 1 0.0121 85.653 -7431.0  
## + baserun\_cs 1 0.0005 85.665 -7430.7  
## + pitching\_hr 1 0.0004 85.665 -7430.7  
## - batting\_hbp\_bi 1 0.1914 85.857 -7429.6  
## - pitching\_bb 1 0.3054 85.971 -7426.6  
## - pitching\_so 1 0.4470 86.112 -7422.8  
## - baserun\_sb 1 0.5464 86.212 -7420.2  
## - batting\_2b 1 0.6689 86.334 -7417.0  
## - batting\_3b 1 0.8611 86.526 -7411.9  
## - batting\_so 1 1.3564 87.022 -7398.9  
## - total\_bases 1 1.9865 87.652 -7382.5  
## - fielding\_dp 1 2.5023 88.168 -7369.1  
## - fielding\_e 1 8.7852 94.450 -7212.5  
##   
## Step: AIC=-7434.25  
## target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + pitching\_h + pitching\_bb + pitching\_so + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + batting\_1B + free\_bases\_num +   
## total\_bases  
##   
## Df Sum of Sq RSS AIC  
## - batting\_bb 1 0.0167 85.697 -7435.8  
## - pitching\_h 1 0.0347 85.716 -7435.3  
## - free\_bases\_num 1 0.0359 85.717 -7435.3  
## <none> 85.681 -7434.2  
## - batting\_1B 1 0.0846 85.765 -7434.0  
## + total\_bases\_allowed 1 0.0156 85.665 -7432.7  
## + batting\_h 1 0.0104 85.670 -7432.5  
## + HR\_over\_OP 1 0.0066 85.674 -7432.4  
## + baserun\_cs 1 0.0015 85.679 -7432.3  
## + pitching\_hr 1 0.0003 85.681 -7432.3  
## - batting\_hbp\_bi 1 0.1936 85.874 -7431.1  
## - pitching\_bb 1 0.2926 85.973 -7428.5  
## - pitching\_so 1 0.4405 86.121 -7424.6  
## - baserun\_sb 1 0.6221 86.303 -7419.8  
## - batting\_3b 1 1.1274 86.808 -7406.5  
## - batting\_2b 1 1.3097 86.991 -7401.7  
## - batting\_so 1 1.3451 87.026 -7400.8  
## - fielding\_dp 1 2.4869 88.168 -7371.1  
## - total\_bases 1 5.3475 91.028 -7298.5  
## - fielding\_e 1 8.8053 94.486 -7213.6  
##   
## Step: AIC=-7435.8  
## target\_wins ~ batting\_2b + batting\_3b + batting\_so + baserun\_sb +   
## pitching\_h + pitching\_bb + pitching\_so + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + batting\_1B + free\_bases\_num + total\_bases  
##   
## Df Sum of Sq RSS AIC  
## - pitching\_h 1 0.0314 85.729 -7437.0  
## <none> 85.697 -7435.8  
## - batting\_1B 1 0.0813 85.779 -7435.6  
## + batting\_bb 1 0.0167 85.681 -7434.2  
## + walks\_over\_OP 1 0.0167 85.681 -7434.2  
## + total\_bases\_allowed 1 0.0145 85.683 -7434.2  
## + batting\_h 1 0.0101 85.687 -7434.1  
## + HR\_over\_OP 1 0.0075 85.690 -7434.0  
## + baserun\_cs 1 0.0014 85.696 -7433.8  
## + pitching\_hr 1 0.0003 85.697 -7433.8  
## - free\_bases\_num 1 0.2108 85.908 -7432.2  
## - pitching\_bb 1 0.3139 86.011 -7429.5  
## - pitching\_so 1 0.4504 86.148 -7425.9  
## - baserun\_sb 1 0.6237 86.321 -7421.3  
## - batting\_3b 1 1.1231 86.821 -7408.2  
## - batting\_2b 1 1.3154 87.013 -7403.1  
## - batting\_so 1 1.3572 87.055 -7402.0  
## - batting\_hbp\_bi 1 1.4961 87.193 -7398.4  
## - fielding\_dp 1 2.4944 88.192 -7372.5  
## - total\_bases 1 5.3330 91.030 -7300.4  
## - fielding\_e 1 8.7892 94.487 -7215.6  
##   
## Step: AIC=-7436.97  
## target\_wins ~ batting\_2b + batting\_3b + batting\_so + baserun\_sb +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + free\_bases\_num + total\_bases  
##   
## Df Sum of Sq RSS AIC  
## - batting\_1B 1 0.0499 85.779 -7437.6  
## <none> 85.729 -7437.0  
## + pitching\_h 1 0.0314 85.697 -7435.8  
## + batting\_h 1 0.0192 85.710 -7435.5  
## + batting\_bb 1 0.0133 85.716 -7435.3  
## + walks\_over\_OP 1 0.0133 85.716 -7435.3  
## + pitching\_hr 1 0.0048 85.724 -7435.1  
## + total\_bases\_allowed 1 0.0021 85.727 -7435.0  
## + HR\_over\_OP 1 0.0016 85.727 -7435.0  
## + baserun\_cs 1 0.0012 85.728 -7435.0  
## - pitching\_so 1 0.4213 86.150 -7427.8  
## - free\_bases\_num 1 0.4948 86.224 -7425.9  
## - pitching\_bb 1 0.5811 86.310 -7423.6  
## - baserun\_sb 1 0.7308 86.460 -7419.7  
## - batting\_3b 1 1.1052 86.834 -7409.8  
## - batting\_so 1 1.3953 87.124 -7402.2  
## - batting\_2b 1 1.5423 87.271 -7398.4  
## - batting\_hbp\_bi 1 1.8417 87.571 -7390.6  
## - fielding\_dp 1 2.4866 88.216 -7373.9  
## - total\_bases 1 7.4968 93.226 -7248.2  
## - fielding\_e 1 9.1865 94.915 -7207.3  
##   
## Step: AIC=-7437.65  
## target\_wins ~ batting\_2b + batting\_3b + batting\_so + baserun\_sb +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## free\_bases\_num + total\_bases  
##   
## Df Sum of Sq RSS AIC  
## <none> 85.779 -7437.6  
## + pitching\_hr 1 0.0531 85.726 -7437.1  
## + batting\_1B 1 0.0499 85.729 -7437.0  
## + batting\_h 1 0.0335 85.745 -7436.5  
## + batting\_bb 1 0.0134 85.765 -7436.0  
## + walks\_over\_OP 1 0.0134 85.765 -7436.0  
## + total\_bases\_allowed 1 0.0049 85.774 -7435.8  
## + baserun\_cs 1 0.0008 85.778 -7435.7  
## + HR\_over\_OP 1 0.0004 85.778 -7435.7  
## + pitching\_h 1 0.0001 85.779 -7435.6  
## - pitching\_so 1 0.4145 86.193 -7428.7  
## - free\_bases\_num 1 0.4704 86.249 -7427.2  
## - pitching\_bb 1 0.5937 86.372 -7423.9  
## - baserun\_sb 1 0.7480 86.527 -7419.9  
## - batting\_3b 1 1.0992 86.878 -7410.7  
## - batting\_so 1 1.5816 87.360 -7398.1  
## - batting\_2b 1 1.6217 87.400 -7397.0  
## - batting\_hbp\_bi 1 1.9217 87.700 -7389.2  
## - fielding\_dp 1 2.4468 88.226 -7375.6  
## - fielding\_e 1 9.1383 94.917 -7209.2  
## - total\_bases 1 9.5063 95.285 -7200.4

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\_so +   
## baserun\_sb + pitching\_bb + pitching\_so + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + free\_bases\_num + total\_bases, data = transformed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.62069 -0.12940 0.00108 0.13000 0.56085   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.001382 0.004415 0.313 0.754362   
## batting\_2b -0.169401 0.025893 -6.542 7.46e-11 \*\*\*  
## batting\_3b 0.160093 0.029723 5.386 7.94e-08 \*\*\*  
## batting\_so -0.548829 0.084946 -6.461 1.27e-10 \*\*\*  
## baserun\_sb 0.121656 0.027380 4.443 9.29e-06 \*\*\*  
## pitching\_bb -0.372319 0.094056 -3.958 7.78e-05 \*\*\*  
## pitching\_so 0.541114 0.163607 3.307 0.000956 \*\*\*  
## fielding\_e -0.527065 0.033938 -15.530 < 2e-16 \*\*\*  
## fielding\_dp -0.171256 0.021311 -8.036 1.48e-15 \*\*\*  
## batting\_hbp\_bi -0.169102 0.023744 -7.122 1.42e-12 \*\*\*  
## free\_bases\_num 0.276863 0.078575 3.524 0.000434 \*\*\*  
## total\_bases 0.550855 0.034776 15.840 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1946 on 2264 degrees of freedom  
## Multiple R-squared: 0.3009, Adjusted R-squared: 0.2975   
## F-statistic: 88.59 on 11 and 2264 DF, p-value: < 2.2e-16

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

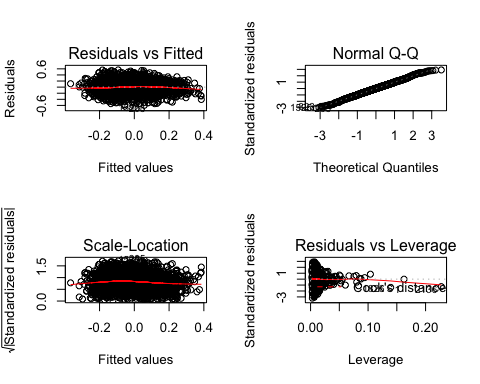
## [1] "MSE equal 0.0376883812165135"

1. Forward direction

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

## Start: AIC=-7425.66  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_h + pitching\_hr +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + free\_bases\_num + 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\_h +   
## pitching\_hr + pitching\_bb + pitching\_so + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + batting\_1B + free\_bases\_num + total\_bases +   
## total\_bases\_allowed + HR\_over\_OP + walks\_over\_OP + SO\_over\_OP,   
## data = transformed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.62430 -0.12992 0.00302 0.12901 0.56473   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.002023 0.004503 0.449 0.653347   
## batting\_h -0.157125 0.218126 -0.720 0.471391   
## batting\_2b -0.154427 0.092172 -1.675 0.093991 .   
## batting\_3b 0.151461 0.082784 1.830 0.067444 .   
## batting\_bb -0.465134 0.724456 -0.642 0.520909   
## batting\_so -0.550649 0.094236 -5.843 5.86e-09 \*\*\*  
## baserun\_sb 0.062377 0.106454 0.586 0.557968   
## baserun\_cs -0.002407 0.037904 -0.064 0.949370   
## pitching\_h -0.122984 0.118058 -1.042 0.297650   
## pitching\_hr -0.105658 0.208575 -0.507 0.612505   
## pitching\_bb -0.367891 0.127904 -2.876 0.004061 \*\*   
## pitching\_so 0.583123 0.175578 3.321 0.000911 \*\*\*  
## fielding\_e -0.526787 0.035344 -14.905 < 2e-16 \*\*\*  
## fielding\_dp -0.174484 0.021579 -8.086 9.96e-16 \*\*\*  
## batting\_hbp\_bi -0.219031 0.098283 -2.229 0.025940 \*   
## batting\_1B 0.133992 0.158747 0.844 0.398725   
## free\_bases\_num 0.660590 0.745334 0.886 0.375550   
## total\_bases 0.729920 0.309296 2.360 0.018363 \*   
## total\_bases\_allowed 0.066528 0.087480 0.760 0.447039   
## HR\_over\_OP -0.051305 0.072371 -0.709 0.478450   
## 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: 0.1948 on 2256 degrees of freedom  
## Multiple R-squared: 0.3021, Adjusted R-squared: 0.2963   
## F-statistic: 51.41 on 19 and 2256 DF, p-value: < 2.2e-16

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

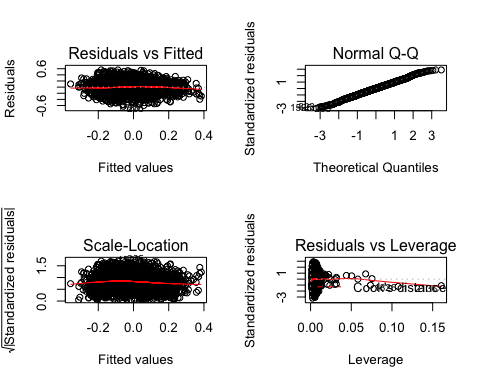
## [1] "MSE equal 0.0376220132488467"

1. Backwards direction

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

## Start: AIC=-7425.66  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_h + pitching\_hr +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + free\_bases\_num + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP + walks\_over\_OP + SO\_over\_OP  
##   
##   
## Step: AIC=-7425.66  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_h + pitching\_hr +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + free\_bases\_num + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP + walks\_over\_OP  
##   
##   
## Step: AIC=-7425.66  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + baserun\_cs + pitching\_h + pitching\_hr +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + free\_bases\_num + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - baserun\_cs 1 0.0002 85.628 -7427.7  
## - pitching\_hr 1 0.0097 85.637 -7427.4  
## - baserun\_sb 1 0.0130 85.641 -7427.3  
## - batting\_bb 1 0.0156 85.643 -7427.2  
## - HR\_over\_OP 1 0.0191 85.647 -7427.2  
## - batting\_h 1 0.0197 85.647 -7427.1  
## - total\_bases\_allowed 1 0.0220 85.650 -7427.1  
## - batting\_1B 1 0.0270 85.655 -7426.9  
## - free\_bases\_num 1 0.0298 85.658 -7426.9  
## - pitching\_h 1 0.0412 85.669 -7426.6  
## <none> 85.628 -7425.7  
## - batting\_2b 1 0.1065 85.734 -7424.8  
## - batting\_3b 1 0.1271 85.755 -7424.3  
## - batting\_hbp\_bi 1 0.1885 85.816 -7422.7  
## - total\_bases 1 0.2114 85.839 -7422.0  
## - pitching\_bb 1 0.3140 85.942 -7419.3  
## - pitching\_so 1 0.4187 86.046 -7416.6  
## - batting\_so 1 1.2960 86.924 -7393.5  
## - fielding\_dp 1 2.4816 88.109 -7362.6  
## - fielding\_e 1 8.4317 94.059 -7213.9  
##   
## Step: AIC=-7427.65  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + pitching\_h + pitching\_hr + pitching\_bb +   
## pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + free\_bases\_num + total\_bases + total\_bases\_allowed +   
## HR\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - pitching\_hr 1 0.0096 85.638 -7429.4  
## - baserun\_sb 1 0.0130 85.641 -7429.3  
## - batting\_bb 1 0.0156 85.643 -7429.2  
## - HR\_over\_OP 1 0.0191 85.647 -7429.1  
## - batting\_h 1 0.0199 85.648 -7429.1  
## - total\_bases\_allowed 1 0.0231 85.651 -7429.0  
## - batting\_1B 1 0.0277 85.656 -7428.9  
## - free\_bases\_num 1 0.0298 85.658 -7428.9  
## - pitching\_h 1 0.0423 85.670 -7428.5  
## <none> 85.628 -7427.7  
## - batting\_2b 1 0.1071 85.735 -7426.8  
## - batting\_3b 1 0.1274 85.755 -7426.3  
## - batting\_hbp\_bi 1 0.1885 85.816 -7424.6  
## - total\_bases 1 0.2113 85.839 -7424.0  
## - pitching\_bb 1 0.3149 85.943 -7421.3  
## - pitching\_so 1 0.4185 86.046 -7418.6  
## - batting\_so 1 1.2974 86.925 -7395.4  
## - fielding\_dp 1 2.4847 88.113 -7364.6  
## - fielding\_e 1 8.4734 94.101 -7214.9  
##   
## Step: AIC=-7429.4  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + pitching\_h + pitching\_bb + pitching\_so +   
## fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B +   
## free\_bases\_num + total\_bases + total\_bases\_allowed + HR\_over\_OP  
##   
## Df Sum of Sq RSS AIC  
## - HR\_over\_OP 1 0.0095 85.647 -7431.1  
## - batting\_h 1 0.0157 85.653 -7431.0  
## - batting\_bb 1 0.0173 85.655 -7430.9  
## - total\_bases\_allowed 1 0.0284 85.666 -7430.6  
## - free\_bases\_num 1 0.0368 85.674 -7430.4  
## - batting\_1B 1 0.0575 85.695 -7429.9  
## - pitching\_h 1 0.0575 85.695 -7429.9  
## <none> 85.638 -7429.4  
## - baserun\_sb 1 0.1819 85.819 -7426.6  
## - batting\_2b 1 0.1856 85.823 -7426.5  
## - batting\_hbp\_bi 1 0.1859 85.823 -7426.5  
## - pitching\_bb 1 0.3081 85.946 -7423.2  
## - pitching\_so 1 0.4138 86.051 -7420.4  
## - total\_bases 1 0.7315 86.369 -7412.0  
## - batting\_3b 1 0.8476 86.485 -7409.0  
## - batting\_so 1 1.3026 86.940 -7397.0  
## - fielding\_dp 1 2.5208 88.158 -7365.4  
## - fielding\_e 1 8.6348 94.272 -7212.8  
##   
## Step: AIC=-7431.14  
## target\_wins ~ batting\_h + batting\_2b + batting\_3b + batting\_bb +   
## batting\_so + baserun\_sb + pitching\_h + pitching\_bb + pitching\_so +   
## fielding\_e + fielding\_dp + batting\_hbp\_bi + batting\_1B +   
## free\_bases\_num + total\_bases + total\_bases\_allowed  
##   
## Df Sum of Sq RSS AIC  
## - batting\_h 1 0.0182 85.665 -7432.7  
## - batting\_bb 1 0.0183 85.665 -7432.7  
## - total\_bases\_allowed 1 0.0234 85.670 -7432.5  
## - free\_bases\_num 1 0.0361 85.683 -7432.2  
## - pitching\_h 1 0.0485 85.695 -7431.9  
## - batting\_1B 1 0.0561 85.703 -7431.7  
## <none> 85.647 -7431.1  
## - baserun\_sb 1 0.1724 85.819 -7428.6  
## - batting\_hbp\_bi 1 0.1893 85.836 -7428.1  
## - batting\_2b 1 0.1951 85.842 -7428.0  
## - pitching\_bb 1 0.3182 85.965 -7424.7  
## - pitching\_so 1 0.4212 86.068 -7422.0  
## - total\_bases 1 0.7882 86.435 -7412.3  
## - batting\_3b 1 0.8382 86.485 -7411.0  
## - batting\_so 1 1.3090 86.956 -7398.6  
## - fielding\_dp 1 2.5113 88.158 -7367.4  
## - fielding\_e 1 8.8032 94.450 -7210.5  
##   
## Step: AIC=-7432.66  
## target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + pitching\_h + pitching\_bb + pitching\_so + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + batting\_1B + free\_bases\_num +   
## total\_bases + total\_bases\_allowed  
##   
## Df Sum of Sq RSS AIC  
## - total\_bases\_allowed 1 0.0156 85.681 -7434.2  
## - batting\_bb 1 0.0177 85.683 -7434.2  
## - free\_bases\_num 1 0.0382 85.703 -7433.6  
## - pitching\_h 1 0.0479 85.713 -7433.4  
## <none> 85.665 -7432.7  
## - batting\_1B 1 0.0930 85.758 -7432.2  
## - batting\_hbp\_bi 1 0.1914 85.857 -7429.6  
## - pitching\_bb 1 0.3054 85.971 -7426.6  
## - pitching\_so 1 0.4470 86.112 -7422.8  
## - baserun\_sb 1 0.5464 86.212 -7420.2  
## - batting\_2b 1 0.6689 86.334 -7417.0  
## - batting\_3b 1 0.8611 86.526 -7411.9  
## - batting\_so 1 1.3564 87.022 -7398.9  
## - total\_bases 1 1.9865 87.652 -7382.5  
## - fielding\_dp 1 2.5023 88.168 -7369.1  
## - fielding\_e 1 8.7852 94.450 -7212.5  
##   
## Step: AIC=-7434.25  
## target\_wins ~ batting\_2b + batting\_3b + batting\_bb + batting\_so +   
## baserun\_sb + pitching\_h + pitching\_bb + pitching\_so + fielding\_e +   
## fielding\_dp + batting\_hbp\_bi + batting\_1B + free\_bases\_num +   
## total\_bases  
##   
## Df Sum of Sq RSS AIC  
## - batting\_bb 1 0.0167 85.697 -7435.8  
## - pitching\_h 1 0.0347 85.716 -7435.3  
## - free\_bases\_num 1 0.0359 85.717 -7435.3  
## <none> 85.681 -7434.2  
## - batting\_1B 1 0.0846 85.765 -7434.0  
## - batting\_hbp\_bi 1 0.1936 85.874 -7431.1  
## - pitching\_bb 1 0.2926 85.973 -7428.5  
## - pitching\_so 1 0.4405 86.121 -7424.6  
## - baserun\_sb 1 0.6221 86.303 -7419.8  
## - batting\_3b 1 1.1274 86.808 -7406.5  
## - batting\_2b 1 1.3097 86.991 -7401.7  
## - batting\_so 1 1.3451 87.026 -7400.8  
## - fielding\_dp 1 2.4869 88.168 -7371.1  
## - total\_bases 1 5.3475 91.028 -7298.5  
## - fielding\_e 1 8.8053 94.486 -7213.6  
##   
## Step: AIC=-7435.8  
## target\_wins ~ batting\_2b + batting\_3b + batting\_so + baserun\_sb +   
## pitching\_h + pitching\_bb + pitching\_so + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + batting\_1B + free\_bases\_num + total\_bases  
##   
## Df Sum of Sq RSS AIC  
## - pitching\_h 1 0.0314 85.729 -7437.0  
## <none> 85.697 -7435.8  
## - batting\_1B 1 0.0813 85.779 -7435.6  
## - free\_bases\_num 1 0.2108 85.908 -7432.2  
## - pitching\_bb 1 0.3139 86.011 -7429.5  
## - pitching\_so 1 0.4504 86.148 -7425.9  
## - baserun\_sb 1 0.6237 86.321 -7421.3  
## - batting\_3b 1 1.1231 86.821 -7408.2  
## - batting\_2b 1 1.3154 87.013 -7403.1  
## - batting\_so 1 1.3572 87.055 -7402.0  
## - batting\_hbp\_bi 1 1.4961 87.193 -7398.4  
## - fielding\_dp 1 2.4944 88.192 -7372.5  
## - total\_bases 1 5.3330 91.030 -7300.4  
## - fielding\_e 1 8.7892 94.487 -7215.6  
##   
## Step: AIC=-7436.97  
## target\_wins ~ batting\_2b + batting\_3b + batting\_so + baserun\_sb +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## batting\_1B + free\_bases\_num + total\_bases  
##   
## Df Sum of Sq RSS AIC  
## - batting\_1B 1 0.0499 85.779 -7437.6  
## <none> 85.729 -7437.0  
## - pitching\_so 1 0.4213 86.150 -7427.8  
## - free\_bases\_num 1 0.4948 86.224 -7425.9  
## - pitching\_bb 1 0.5811 86.310 -7423.6  
## - baserun\_sb 1 0.7308 86.460 -7419.7  
## - batting\_3b 1 1.1052 86.834 -7409.8  
## - batting\_so 1 1.3953 87.124 -7402.2  
## - batting\_2b 1 1.5423 87.271 -7398.4  
## - batting\_hbp\_bi 1 1.8417 87.571 -7390.6  
## - fielding\_dp 1 2.4866 88.216 -7373.9  
## - total\_bases 1 7.4968 93.226 -7248.2  
## - fielding\_e 1 9.1865 94.915 -7207.3  
##   
## Step: AIC=-7437.65  
## target\_wins ~ batting\_2b + batting\_3b + batting\_so + baserun\_sb +   
## pitching\_bb + pitching\_so + fielding\_e + fielding\_dp + batting\_hbp\_bi +   
## free\_bases\_num + total\_bases  
##   
## Df Sum of Sq RSS AIC  
## <none> 85.779 -7437.6  
## - pitching\_so 1 0.4145 86.193 -7428.7  
## - free\_bases\_num 1 0.4704 86.249 -7427.2  
## - pitching\_bb 1 0.5937 86.372 -7423.9  
## - baserun\_sb 1 0.7480 86.527 -7419.9  
## - batting\_3b 1 1.0992 86.878 -7410.7  
## - batting\_so 1 1.5816 87.360 -7398.1  
## - batting\_2b 1 1.6217 87.400 -7397.0  
## - batting\_hbp\_bi 1 1.9217 87.700 -7389.2  
## - fielding\_dp 1 2.4468 88.226 -7375.6  
## - fielding\_e 1 9.1383 94.917 -7209.2  
## - total\_bases 1 9.5063 95.285 -7200.4

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\_so +   
## baserun\_sb + pitching\_bb + pitching\_so + fielding\_e + fielding\_dp +   
## batting\_hbp\_bi + free\_bases\_num + total\_bases, data = transformed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.62069 -0.12940 0.00108 0.13000 0.56085   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.001382 0.004415 0.313 0.754362   
## batting\_2b -0.169401 0.025893 -6.542 7.46e-11 \*\*\*  
## batting\_3b 0.160093 0.029723 5.386 7.94e-08 \*\*\*  
## batting\_so -0.548829 0.084946 -6.461 1.27e-10 \*\*\*  
## baserun\_sb 0.121656 0.027380 4.443 9.29e-06 \*\*\*  
## pitching\_bb -0.372319 0.094056 -3.958 7.78e-05 \*\*\*  
## pitching\_so 0.541114 0.163607 3.307 0.000956 \*\*\*  
## fielding\_e -0.527065 0.033938 -15.530 < 2e-16 \*\*\*  
## fielding\_dp -0.171256 0.021311 -8.036 1.48e-15 \*\*\*  
## batting\_hbp\_bi -0.169102 0.023744 -7.122 1.42e-12 \*\*\*  
## free\_bases\_num 0.276863 0.078575 3.524 0.000434 \*\*\*  
## total\_bases 0.550855 0.034776 15.840 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1946 on 2264 degrees of freedom  
## Multiple R-squared: 0.3009, Adjusted R-squared: 0.2975   
## F-statistic: 88.59 on 11 and 2264 DF, p-value: < 2.2e-16

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

## [1] "MSE equal 0.0376883812165135"

## Conclusion

It definitely made a difference when the transformation were applied. One can see the difference in the residual plots. The residual is now normal(per QQ plot), and there are no patterns when we look at he Rsiduals Vs Fitted plot. When looking at the Rsquared and Adjusted Rsquared together with the residual plots, it's easy to conclude that the model with the stepwise approach together with the transformations is the one that leads to a better model.

Though RMSE and Rsquared from the other models seem to suggest otherwise, the stepwise model appears to be more stable. I also noticed by looking at the Cook's Distance plot that there are influncial observations, but for some reason I could not get robust regression to work. From my understanding, robust regression would put less enphasis on those data points, leading to a more accurate model.