Homework #1: Baseball Analysis

Data 621 Business Analytics and Data Mining

Aadi Kalloo, Nathan Lim, Asher Meyers, Daniel Smilowitz, Logan Thomson

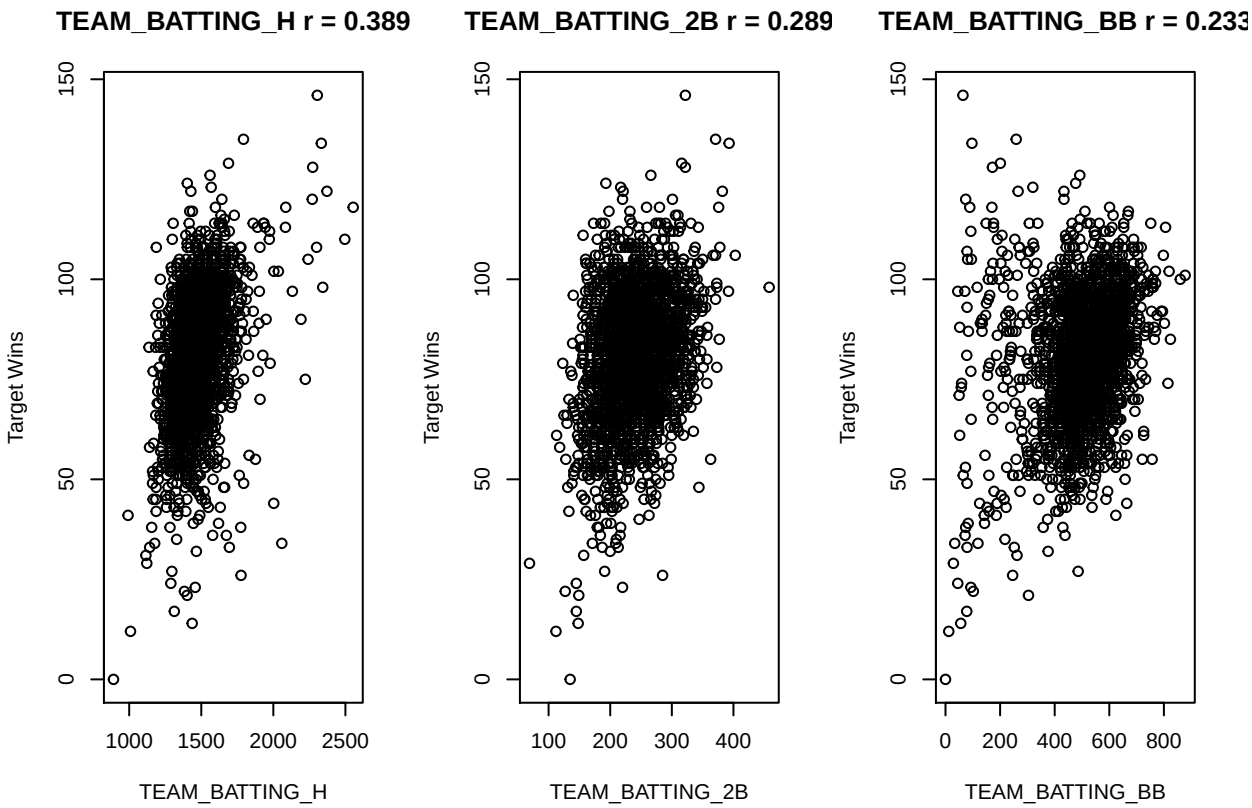
Due June 19, 2016

Data Exploration

The data analyzed in this report includes 2276 professional baseball teams for the years 1871-2006. In total, 16 variables were present in the data provided. Included below is a summary of descriptive statistics, correlations to wins, and the number of missing values for each variable in the provided data set:

	VAR_NAME	MEAN	MEDIAN	CORRELATION TO WINS (r)	NUM_MISSING
2	TARGET_WINS	80.79086	82.0	NA	NA
1	TEAM_BASERUN_CS	52.80386	49.0	0.0224041	772
21	TEAM_BASERUN_SB	124.76177	101.0	0.1351389	131
3	TEAM_BATTING_2B	241.24692	238.0	0.2891036	0
4	TEAM_BATTING_3B	55.25000	47.0	0.1426084	0
5	TEAM_BATTING_BB	501.55888	512.0	0.2325599	0
6	TEAM_BATTING_H	1469.26977	1454.0	0.3887675	0
7	TEAM_BATTING_HBP	59.35602	58.0	0.0735042	2085
8	TEAM_BATTING_HR	99.61204	102.0	0.1761532	0
9	TEAM_BATTING_SO	735.60534	750.0	-0.0317507	102
10	TEAM_FIELDING_DP	146.38794	149.0	-0.0348506	286
11	TEAM_FIELDING_E	246.48067	159.0	-0.1764848	0
12	TEAM_PITCHING_BB	553.00791	536.5	0.1241745	0
13	TEAM_PITCHING_H	1779.21046	1518.0	-0.1099371	0
14	TEAM_PITCHING_HR	105.69859	107.0	0.1890137	0
15	TEAM_PITCHING_SO	817.73045	813.5	-0.0784361	102

Below are graphs that show the relationship to *Target Wins* for the three variables with the highest correlation coefficient:



The full array of correlations graphs may be found in Appendix A.

Data Preparation

It was determined that the *Hits By Pitch* variable had too many missing values to be useful for regression, and thus this variable was excluded from the model building process. As shown in Table 1 above, there are several variables that have missing values. The attempted solution to this problem involved imputation using the median for each variable in the data set. A summary of the data is shown here again for inspection and confirmation of similarity between the old and new data sets:

Missing Values Imputed With Median

	VAR_NAME	MEAN	MEDIAN	CORRELATION TO WINS (r)	NUM_MISSING
2	TEAM_BATTING_H	1469.26977	1454.0	NA	NA
1	TEAM_BASERUN_CS	51.51362	49.0	0.0159598	0
21	TEAM_BASERUN_SB	123.39411	101.0	0.1236109	0
3	TEAM_BATTING_2B	241.24692	238.0	0.2891036	0
4	TEAM_BATTING_3B	55.25000	47.0	0.1426084	0
5	TEAM_BATTING_BB	501.55888	512.0	0.2325599	0
6	TEAM_BATTING_HBP	58.11380	58.0	0.0165164	0
7	TEAM_BATTING_HR	99.61204	102.0	0.1761532	0
8	TEAM_BATTING_SO	736.25044	750.0	-0.0305814	0
9	TEAM_FIELDING_DP	146.71617	149.0	-0.0300863	0
10	TEAM_FIELDING_E	246.48067	159.0	-0.1764848	0
11	TEAM_PITCHING_BB	553.00791	536.5	0.1241745	0
12	TEAM_PITCHING_H	1779.21046	1518.0	-0.1099371	0
13	TEAM_PITCHING_HR	105.69859	107.0	0.1890137	0
14	TEAM_PITCHING_SO	817.54086	813.5	-0.0757997	0

The dataset contains 17 columns - an index column (INDEX), a response column (TARGET_WINS) and 15 predictor columns. There are 2,276 observations - but there are many missing values for many of the predictors.

Two predictors in particular stand out:

	Predictor Name	Description	Impact	% Missing	r with Response	p-Value
a	TEAM_BATTING_HBP	Batters hit by pitch (free base)	Positive	91.6%	0.07	0.31
b	TEAM_BASERUN_CS	Strikeouts by batters	Negative	33.9%	0.02	0.39

Including these predictors in our dataset would mean that we would either have to a) forgo a significant chunk of our data (34% or 92%) or b) impute a large number of data points. Their correlation coefficients with the response are less than an absolute value of 7%; the p values of a simple one variable linear regression using them and the response yields models of no statistical significance (i.e. $p > 0.05$). Thus, it seems safe to exclude these predictors from our models. This way, we avoid the twin pitfalls of mass exclusion and imputation.

Further exclusions to the data were made:

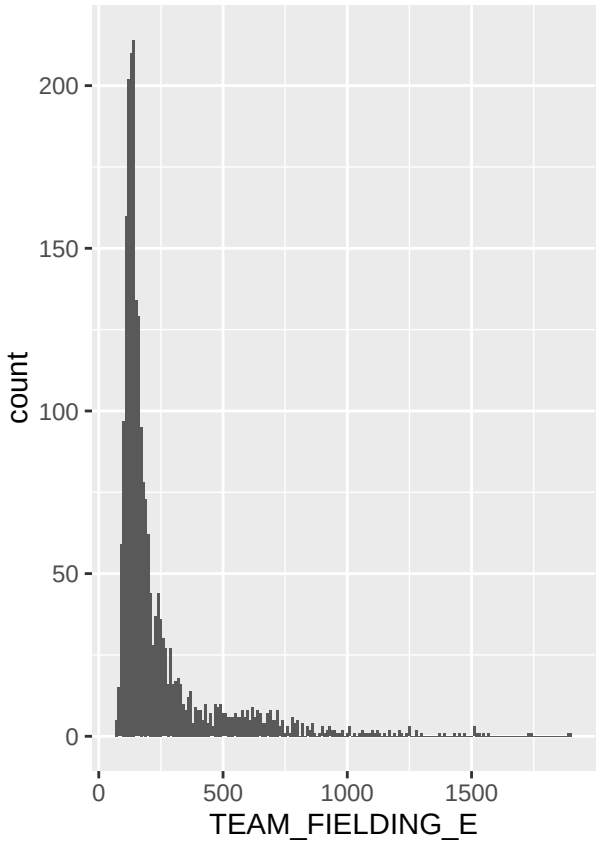
Exclusion	Explanation
INDEX == 1347	This row had a suspicious set of zero entries
TEAM_BATTING_BB == 0	Anomalously low walk count (expected occurences of a zero value for this predictor are zero)
TEAM_BATTING_SO	Outside of recognized records link
TEAM_BATTING_HR	Outside of recognized records link

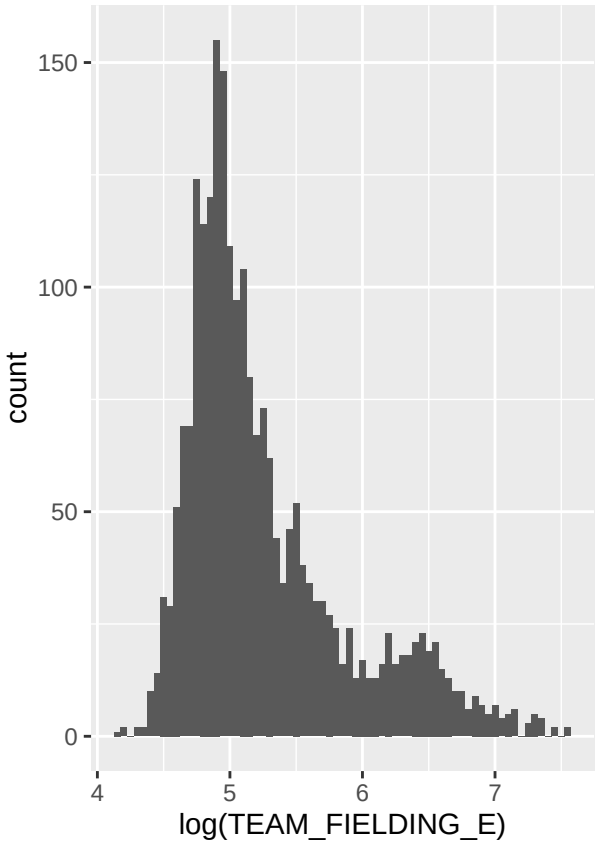
It should be noted that the records excluded from the first two rows of the table above are the same exact points (which would technically make the second exclusion redundant...). That suggests that for whatever reason, strikeouts were not recorded for those rows, but were marked as zero. Those two predictors have the same number of NA values, 102, suggesting their recording method was linked somehow.

Model Creation

Model –(Nathan)

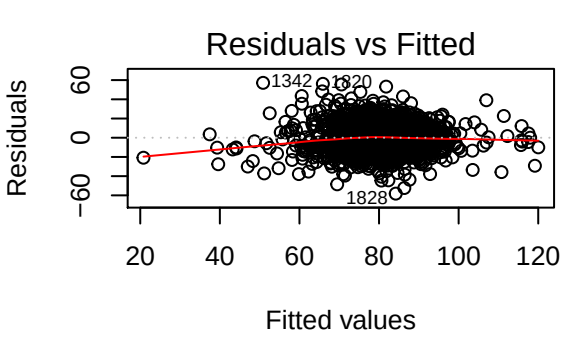
```
g1 <- ggplot(df_new, aes(x=TEAM_FIELDING_E)) + geom_histogram(binwidth = 10)
g2 <- ggplot(df_new, aes(x=log(TEAM_FIELDING_E))) + geom_histogram(binwidth = 0.05)
grid.arrange(g1, g2, ncol=2)
```

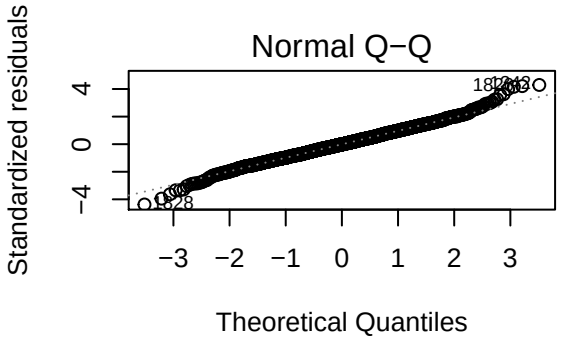


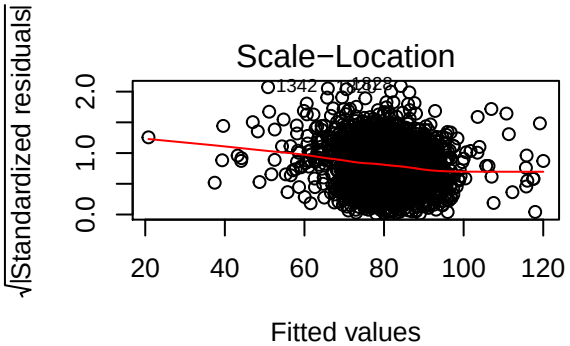


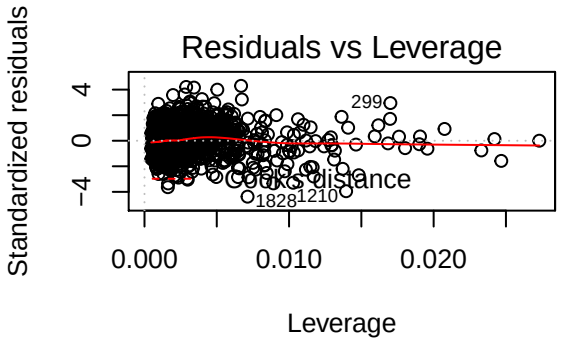

```
model2<- lm(TARGET_WINS ~  TEAM_BATTING_H + TEAM_BASERUN_SB + TEAM_FIELDING_DP +
              log(TEAM_FIELDING_E), df_new)

par(mfrow=c(2,2)); plot(model2)
```

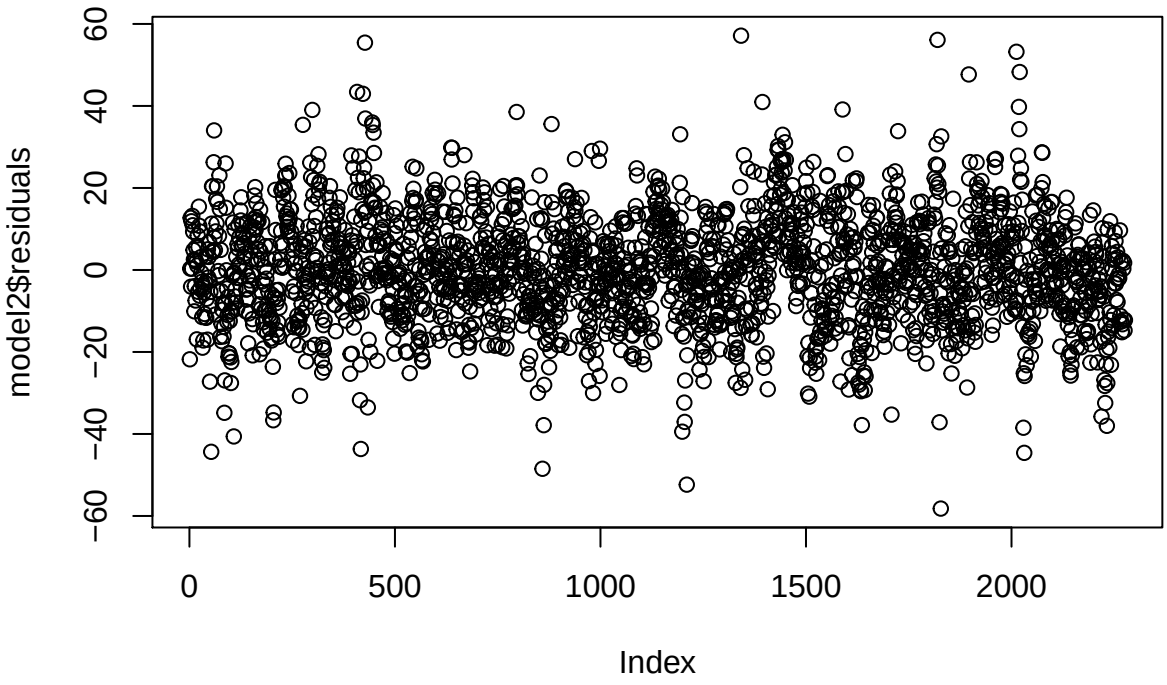









```
par(mfrow=c(1,1)); plot(model2$residuals)
```

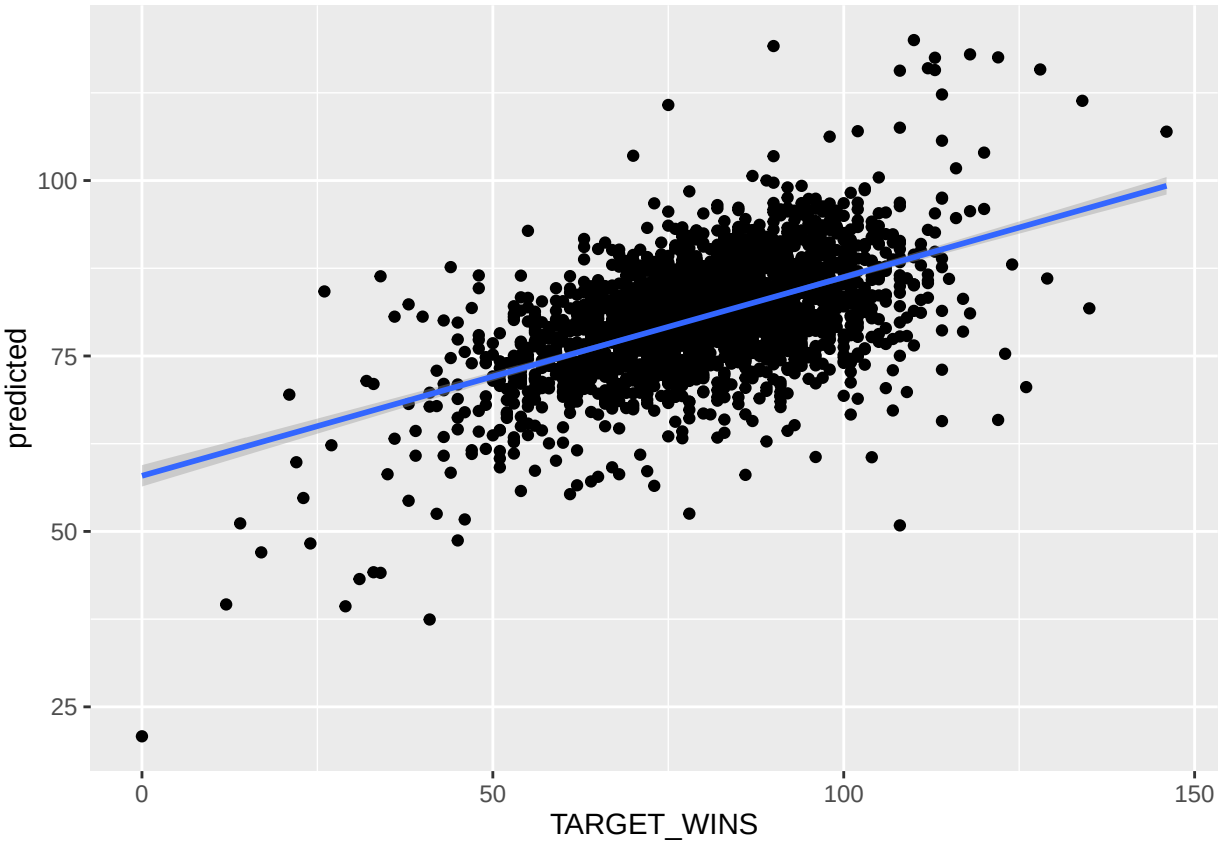



```
summary(model2)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BASERUN_SB +
##     TEAM_FIELDING_DP + log(TEAM_FIELDING_E), data = df_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.197  -8.922  -0.121   8.638  57.139
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           69.755250   3.989775  17.484   <2e-16 ***
## TEAM_BATTING_H         0.052937   0.002044  25.896   <2e-16 ***
## TEAM_BASERUN_SB        0.039473   0.003715  10.625   <2e-16 ***
## TEAM_FIELDING_DP      -0.105382   0.012468  -8.453   <2e-16 ***
## log(TEAM_FIELDING_E) -10.658801   0.542799 -19.637   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.35 on 2271 degrees of freedom
## Multiple R-squared:  0.283,  Adjusted R-squared:  0.2817
## F-statistic: 224.1 on 4 and 2271 DF,  p-value: < 2.2e-16
```

```
df_new$residuals <- model2$residuals
df_new$predicted <- model2$fitted.values

ggplot(df_new, aes(x=TARGET_WINS, y=predicted)) +
  geom_point() + stat_smooth(method="lm")
```



Model 1

Description:

Relevant code for checking correlation coefficients and p values:

```
#dfraw <- read.csv(url("https://raw.githubusercontent.com/dsmilo/DATA621/master/HW1/data/moneyball-training-da
dfraw = trainingdata
dfHBP <- dfraw[!is.na(dfraw$TEAM_BATTING_HBP),] #Create df without null values for TEAM_BATTING_HBP
paste0("correlation coefficient between response and TEAM_BATTING_HBP: ", cor(dfHBP$TARGET_WINS,dfHBP$TEAM_BAT
```

```
## [1] "correlation coefficient between response and TEAM_BATTING_HBP: 0.0165164113768568"
```

```
summary(lm(TARGET_WINS~TEAM_BATTING_HBP, dfHBP))#See summary of linear regression model using TEAM_BATTING_HBP
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_HBP, data = dfHBP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.783  -9.783   1.217  11.217  65.217
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      76.77638    5.10703  15.033   <2e-16 ***
## TEAM_BATTING_HBP  0.06908    0.08770   0.788    0.431
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.75 on 2274 degrees of freedom
## Multiple R-squared:  0.0002728,  Adjusted R-squared:  -0.0001668
## F-statistic: 0.6205 on 1 and 2274 DF,  p-value: 0.4309
```

```r
dfCS <- dfraw[!is.na(dfraw$TEAM_BASERUN_CS),]#Create df without null values for TEAM_BASERUN_CS
paste0("correlation coefficient between response and TEAM_BASERUN_CS: ", cor(dfCS$TARGET_WINS,dfCS$TEAM_BASERU
```

```
## [1] "correlation coefficient between response and TEAM_BASERUN_CS: 0.0159598171918147"
```

```r
summary(lm(TARGET_WINS~TEAM_BASERUN_CS, dfCS))#See summary of linear regression model using TEAM_BASERUN_CS
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BASERUN_CS, data = dfCS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.100  -9.677   1.203  10.978  65.243
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     80.10001    0.96583  82.934   <2e-16 ***
## TEAM_BASERUN_CS  0.01341    0.01762   0.761    0.447
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.75 on 2274 degrees of freedom
## Multiple R-squared:  0.0002547,  Adjusted R-squared:  -0.0001849
## F-statistic: 0.5794 on 1 and 2274 DF,  p-value: 0.4466
```

I then created a linear regression, and created additional, improved regression models by removing predictors with low significance, until the f-statistic of the regression stopped increasing with the removal of predictors. [The f-stats mentioned in the comments may have changed]

```r
#dfraw <- read.csv(url("https://raw.githubusercontent.com/dsmilo/DATA621/master/HW1/data/moneyball-training-da
dfraw = trainingdata_bk
dfremove <- subset(dfraw, INDEX == 1347 | TEAM_BATTING_BB == 0 |
                     TEAM_BATTING_3B < 11 | TEAM_BATTING_3B > 153 | # http://www.baseball-almanac.com/rb_trip2
                     TEAM_BATTING_HR < 3 | TEAM_BATTING_HR > 264 |#http://www.baseball-almanac.com/recbooks/rb
                     TEAM_PITCHING_SO > 1781 | #http://www.baseball-almanac.com/recbooks/rb_strik.shtml
                     TEAM_BATTING_SO < 308 | TEAM_BATTING_SO > 1535 #http://www.baseball-almanac.com/recbooks/
                   )$INDEX
#length(dfremove)
df <- subset(dfraw, !(INDEX %in% dfremove))
#str(df)
df <- df[, -c(1,10,11,15)] #Remove caught stealing and hit by pitcher variables, and pitching strikeouts.
#View(df)
#View(df1)
#summary(df)
#str(df)

fit <- lm(TARGET_WINS~.,df)
summary(fit)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.245  -7.289   0.086   6.892  29.631
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      58.863622   6.036992   9.750  < 2e-16 ***
## TEAM_BATTING_H   -0.016747   0.015074  -1.111  0.26671
## TEAM_BATTING_2B  -0.049065   0.008885  -5.522 3.83e-08 ***
## TEAM_BATTING_3B   0.183163   0.019009   9.635  < 2e-16 ***
## TEAM_BATTING_HR   0.252620   0.064723   3.903 9.84e-05 ***
## TEAM_BATTING_BB   0.125582   0.042218   2.975  0.00297 **
## TEAM_BATTING_SO  -0.022550   0.002306  -9.781  < 2e-16 ***
## TEAM_BASERUN_SB   0.069681   0.005543  12.571  < 2e-16 ***
## TEAM_PITCHING_H   0.043626   0.013602   3.207  0.00136 **
## TEAM_PITCHING_HR -0.146839   0.061752  -2.378  0.01751 *
## TEAM_PITCHING_BB -0.087159   0.040135  -2.172  0.03001 *
## TEAM_FIELDING_E  -0.118497   0.007147 -16.580  < 2e-16 ***
## TEAM_FIELDING_DP -0.112560   0.012292  -9.157  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.19 on 1822 degrees of freedom
##   (364 observations deleted due to missingness)
## Multiple R-squared:  0.4043, Adjusted R-squared:  0.4003
## F-statistic:   103 on 12 and 1822 DF,  p-value: < 2.2e-16
```

```
fit1 <- update(fit, .~.-TEAM_BATTING_H)
summary(fit1)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_BB + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.412  -7.233   0.053   6.934  29.676
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      57.006825   5.801417   9.826  < 2e-16 ***
## TEAM_BATTING_2B  -0.051344   0.008646  -5.939 3.43e-09 ***
## TEAM_BATTING_3B   0.179945   0.018789   9.577  < 2e-16 ***
## TEAM_BATTING_HR   0.254991   0.064691   3.942 8.40e-05 ***
## TEAM_BATTING_BB   0.084050   0.019622   4.283 1.94e-05 ***
## TEAM_BATTING_SO  -0.022139   0.002276  -9.728  < 2e-16 ***
## TEAM_BASERUN_SB   0.069844   0.005541  12.604  < 2e-16 ***
## TEAM_PITCHING_H   0.029152   0.003910   7.455 1.38e-13 ***
## TEAM_PITCHING_HR -0.150208   0.061681  -2.435  0.01498 *
## TEAM_PITCHING_BB -0.047551   0.018435  -2.579  0.00998 **
## TEAM_FIELDING_E  -0.116912   0.007004 -16.693  < 2e-16 ***
## TEAM_FIELDING_DP -0.113484   0.012264  -9.253  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.19 on 1823 degrees of freedom
##   (364 observations deleted due to missingness)
## Multiple R-squared:  0.4039, Adjusted R-squared:  0.4003
## F-statistic: 112.3 on 11 and 1823 DF,  p-value: < 2.2e-16
```

```r
fit2 <- update(fit1, .~.-TEAM_PITCHING_HR)
summary(fit2)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_PITCHING_H + TEAM_PITCHING_BB + TEAM_FIELDING_E + TEAM_FIELDING_DP,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.560  -7.244   0.085   6.983  29.632
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      57.424391   5.806715   9.889  < 2e-16 ***
## TEAM_BATTING_2B  -0.051190   0.008657  -5.913  4.0e-09 ***
## TEAM_BATTING_3B   0.178558   0.018805   9.495  < 2e-16 ***
## TEAM_BATTING_HR   0.099023   0.009123  10.854  < 2e-16 ***
## TEAM_BATTING_BB   0.119818   0.013029   9.196  < 2e-16 ***
## TEAM_BATTING_SO  -0.022404   0.002276  -9.842  < 2e-16 ***
## TEAM_BASERUN_SB   0.070701   0.005538  12.767  < 2e-16 ***
## TEAM_PITCHING_H   0.029020   0.003915   7.412  1.9e-13 ***
## TEAM_PITCHING_BB -0.082098   0.011789  -6.964  4.6e-12 ***
## TEAM_FIELDING_E  -0.114487   0.006942 -16.492  < 2e-16 ***
## TEAM_FIELDING_DP -0.115132   0.012262  -9.389  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.2 on 1824 degrees of freedom
##   (364 observations deleted due to missingness)
## Multiple R-squared:  0.4019, Adjusted R-squared:  0.3986
## F-statistic: 122.6 on 10 and 1824 DF,  p-value: < 2.2e-16
```

```r
fit3 <- update(fit2, .~.-TEAM_BATTING_2B)
summary(fit3) #F stat of 130
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_3B + TEAM_BATTING_HR +
##     TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_PITCHING_H +
##     TEAM_PITCHING_BB + TEAM_FIELDING_E + TEAM_FIELDING_DP, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.429  -7.344   0.018   7.050  29.434
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      69.592601   5.480122  12.699  < 2e-16 ***
## TEAM_BATTING_3B   0.185085   0.018947   9.769  < 2e-16 ***
## TEAM_BATTING_HR   0.105225   0.009147  11.504  < 2e-16 ***
## TEAM_BATTING_BB   0.074679   0.010657   7.008 3.40e-12 ***
## TEAM_BATTING_SO  -0.025897   0.002219 -11.672  < 2e-16 ***
## TEAM_BASERUN_SB   0.072748   0.005578  13.042  < 2e-16 ***
## TEAM_PITCHING_H   0.013295   0.002900   4.584 4.87e-06 ***
## TEAM_PITCHING_BB -0.040961   0.009605  -4.264 2.11e-05 ***
```

```
## TEAM_FIELDING_E  -0.106346   0.006867 -15.487  < 2e-16 ***
## TEAM_FIELDING_DP -0.111995   0.012364  -9.058  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.3 on 1825 degrees of freedom
##   (364 observations deleted due to missingness)
## Multiple R-squared:  0.3905, Adjusted R-squared:  0.3875
## F-statistic: 129.9 on 9 and 1825 DF,  p-value: < 2.2e-16
```

```
fit4 <- update(fit3, .~.-TEAM_PITCHING_BB)
summary(fit4)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_3B + TEAM_BATTING_HR +
##     TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_PITCHING_H +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.772  -7.276   0.304   7.032  30.376
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      83.723140   4.385328  19.092   <2e-16 ***
## TEAM_BATTING_3B   0.210286   0.018086  11.627   <2e-16 ***
## TEAM_BATTING_HR   0.123072   0.008171  15.061   <2e-16 ***
## TEAM_BATTING_BB   0.031259   0.003160   9.893   <2e-16 ***
## TEAM_BATTING_SO  -0.028750   0.002125 -13.528   <2e-16 ***
## TEAM_BASERUN_SB   0.074797   0.005583  13.396   <2e-16 ***
## TEAM_PITCHING_H   0.003197   0.001682   1.900   0.0576 .
## TEAM_FIELDING_E  -0.107843   0.006890 -15.652   <2e-16 ***
## TEAM_FIELDING_DP -0.108050   0.012387  -8.723   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.35 on 1826 degrees of freedom
##   (364 observations deleted due to missingness)
## Multiple R-squared:  0.3844, Adjusted R-squared:  0.3817
## F-statistic: 142.5 on 8 and 1826 DF,  p-value: < 2.2e-16
```

```
fit5 <- update(fit4, .~.-TEAM_PITCHING_H)
summary(fit5)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_3B + TEAM_BATTING_HR +
##     TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.586  -7.300   0.315   6.923  30.970
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      88.305662   3.665281  24.092   <2e-16 ***
## TEAM_BATTING_3B   0.217968   0.017641  12.356   <2e-16 ***
```

```
## TEAM_BATTING_HR   0.128691   0.007623  16.883   <2e-16 ***
## TEAM_BATTING_BB   0.031159   0.003162   9.855   <2e-16 ***
## TEAM_BATTING_SO  -0.030171   0.001991 -15.155   <2e-16 ***
## TEAM_BASERUN_SB   0.076368   0.005526  13.820   <2e-16 ***
## TEAM_FIELDING_E  -0.107336   0.006890 -15.579   <2e-16 ***
## TEAM_FIELDING_DP -0.106716   0.012376  -8.623   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.35 on 1827 degrees of freedom
##   (364 observations deleted due to missingness)
## Multiple R-squared:  0.3832, Adjusted R-squared:  0.3808
## F-statistic: 162.1 on 7 and 1827 DF,  p-value: < 2.2e-16
```

```
fit6 <- update(fit5, .~.-TEAM_FIELDING_DP) #Wrong sign on predictor Fielding
summary(fit6)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_3B + TEAM_BATTING_HR +
##     TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_FIELDING_E,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.414  -7.783   0.165   7.539  44.648
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     65.387690   2.509403  26.057   <2e-16 ***
## TEAM_BATTING_3B  0.198896   0.017115  11.621   <2e-16 ***
## TEAM_BATTING_HR  0.120484   0.007464  16.141   <2e-16 ***
## TEAM_BATTING_BB  0.028110   0.003116   9.021   <2e-16 ***
## TEAM_BATTING_SO -0.024350   0.001939 -12.557   <2e-16 ***
## TEAM_BASERUN_SB  0.084101   0.004606  18.258   <2e-16 ***
## TEAM_FIELDING_E -0.075656   0.003913 -19.337   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.14 on 1994 degrees of freedom
##   (198 observations deleted due to missingness)
## Multiple R-squared:  0.3486, Adjusted R-squared:  0.3467
## F-statistic: 177.9 on 6 and 1994 DF,  p-value: < 2.2e-16
```

```
#Correlation Matrix
#View(round(cor(df), 2))

#These are variables that I tried but didn't turn out to be valuable
df$TEAM_BATTING_1B <- df$TEAM_BATTING_H - df$TEAM_BATTING_2B - df$TEAM_BATTING_3B - df$TEAM_BATTING_HR #Single
df$TEAM_BATTING_HRP <- df$TEAM_BATTING_HR/df$TEAM_BATTING_H #Home runs as a percentage of base hits
df$TEAM_BATTING_HSO <- df$TEAM_BATTING_H/df$TEAM_BATTING_SO #Ratio of hits to strikeouts
```

Create a linear model using all predictors. The INDEX column is excluded.

```
FullModel <- lm(TARGET_WINS ~.-INDEX, trainingDataRaw)
summary(FullModel) #Summary of full model
```

```
##
```

```
## Call:
## lm(formula = TARGET_WINS ~ . - INDEX, data = trainingDataRaw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8708  -5.6564  -0.0599   5.2545  22.9274
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      60.28826   19.67842   3.064  0.00253 **
## TEAM_BATTING_H    1.91348    2.76139   0.693  0.48927
## TEAM_BATTING_2B   0.02639    0.03029   0.871  0.38484
## TEAM_BATTING_3B  -0.10118    0.07751  -1.305  0.19348
## TEAM_BATTING_HR  -4.84371   10.50851  -0.461  0.64542
## TEAM_BATTING_BB  -4.45969    3.63624  -1.226  0.22167
## TEAM_BATTING_SO   0.34196    2.59876   0.132  0.89546
## TEAM_BASERUN_SB   0.03304    0.02867   1.152  0.25071
## TEAM_BASERUN_CS  -0.01104    0.07143  -0.155  0.87730
## TEAM_BATTING_HBP  0.08247    0.04960   1.663  0.09815 .
## TEAM_PITCHING_H  -1.89096    2.76095  -0.685  0.49432
## TEAM_PITCHING_HR  4.93043   10.50664   0.469  0.63946
## TEAM_PITCHING_BB  4.51089    3.63372   1.241  0.21612
## TEAM_PITCHING_SO -0.37364    2.59705  -0.144  0.88577
## TEAM_FIELDING_E  -0.17204    0.04140  -4.155 5.08e-05 ***
## TEAM_FIELDING_DP -0.10819    0.03654  -2.961  0.00349 **
## ---
## 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
```

Put full model through stepwise regression, where predictors with less significance are sequentially removed.

```
stepFull <- step(FullModel)
```

```
## Start:  AIC=831.31
## TARGET_WINS ~ (INDEX + TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_BASERUN_CS + TEAM_BATTING_HBP + TEAM_PITCHING_H + TEAM_PITCHING_HR +
##     TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP) -
##     INDEX
##
##                     Df Sum of Sq   RSS    AIC
## - TEAM_BATTING_SO    1      1.24 12547 829.33
## - TEAM_PITCHING_SO   1      1.48 12547 829.33
## - TEAM_BASERUN_CS    1      1.71 12548 829.34
## - TEAM_BATTING_HR    1     15.23 12561 829.54
## - TEAM_PITCHING_HR   1     15.79 12562 829.55
## - TEAM_PITCHING_H    1     33.63 12580 829.82
## - TEAM_BATTING_H     1     34.42 12580 829.83
## - TEAM_BATTING_2B    1     54.41 12600 830.14
## - TEAM_BASERUN_SB    1     95.22 12641 830.76
## - TEAM_BATTING_BB    1    107.84 12654 830.95
## - TEAM_PITCHING_BB   1    110.48 12656 830.99
## - TEAM_BATTING_3B    1    122.16 12668 831.16
## <none>                          12546 831.31
```

```
## - TEAM_BATTING_HBP  1    198.21 12744 832.31
## - TEAM_FIELDING_DP  1    628.49 13174 838.65
## - TEAM_FIELDING_E   1   1237.79 13784 847.28
##
## Step:  AIC=829.33
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BASERUN_SB + TEAM_BASERUN_CS +
##     TEAM_BATTING_HBP + TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_BB +
##     TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq   RSS    AIC
## - TEAM_BASERUN_CS   1      1.59 12549 827.35
## - TEAM_BATTING_HR   1     15.82 12563 827.57
## - TEAM_PITCHING_HR  1     16.39 12564 827.58
## - TEAM_BATTING_2B   1     53.47 12601 828.14
## - TEAM_PITCHING_H   1     88.45 12636 828.67
## - TEAM_BATTING_H    1     90.30 12637 828.70
## - TEAM_BASERUN_SB   1     94.19 12641 828.76
## - TEAM_BATTING_BB   1    107.95 12655 828.97
## - TEAM_PITCHING_BB  1    110.60 12658 829.01
## - TEAM_BATTING_3B   1    122.20 12669 829.18
## <none>                          12547 829.33
## - TEAM_BATTING_HBP  1    197.11 12744 830.31
## - TEAM_FIELDING_DP  1    630.68 13178 836.70
## - TEAM_FIELDING_E   1   1240.80 13788 845.34
## - TEAM_PITCHING_SO  1   1312.89 13860 846.34
##
## Step:  AIC=827.35
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BASERUN_SB + TEAM_BATTING_HBP +
##     TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_BB + TEAM_PITCHING_SO +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq   RSS    AIC
## - TEAM_BATTING_HR   1     16.06 12565 825.60
## - TEAM_PITCHING_HR  1     16.64 12565 825.61
## - TEAM_BATTING_2B   1     53.05 12602 826.16
## - TEAM_PITCHING_H   1     90.24 12639 826.72
## - TEAM_BATTING_H    1     92.13 12641 826.75
## - TEAM_BATTING_BB   1    110.31 12659 827.03
## - TEAM_PITCHING_BB  1    113.00 12662 827.07
## - TEAM_BASERUN_SB   1    123.42 12672 827.22
## - TEAM_BATTING_3B   1    129.33 12678 827.31
## <none>                          12549 827.35
## - TEAM_BATTING_HBP  1    197.23 12746 828.33
## - TEAM_FIELDING_DP  1    635.62 13184 834.79
## - TEAM_PITCHING_SO  1   1311.88 13861 844.35
## - TEAM_FIELDING_E   1   1322.05 13871 844.49
##
## Step:  AIC=825.6
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_BB + TEAM_BASERUN_SB + TEAM_BATTING_HBP + TEAM_PITCHING_H +
##     TEAM_PITCHING_HR + TEAM_PITCHING_BB + TEAM_PITCHING_SO +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq   RSS    AIC
## - TEAM_BATTING_2B   1     55.48 12620 824.44
## - TEAM_PITCHING_H   1     89.26 12654 824.95
## - TEAM_BATTING_H    1     91.97 12657 824.99
## - TEAM_BATTING_BB   1    104.58 12669 825.18
```

```
## - TEAM_PITCHING_BB  1    107.19 12672 825.22
## <none>                        12565 825.60
## - TEAM_BATTING_3B   1    137.48 12702 825.68
## - TEAM_BASERUN_SB   1    146.90 12712 825.82
## - TEAM_BATTING_HBP  1    200.36 12765 826.62
## - TEAM_FIELDING_DP  1    628.95 13194 832.93
## - TEAM_PITCHING_HR  1    853.54 13418 836.15
## - TEAM_PITCHING_SO  1   1316.68 13882 842.63
## - TEAM_FIELDING_E   1   1333.15 13898 842.86
##
## Step:  AIC=824.44
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_3B + TEAM_BATTING_BB +
##     TEAM_BASERUN_SB + TEAM_BATTING_HBP + TEAM_PITCHING_H + TEAM_PITCHING_HR +
##     TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq   RSS    AIC
## - TEAM_PITCHING_H   1     84.47 12705 823.71
## - TEAM_BATTING_H    1     87.79 12708 823.76
## - TEAM_BATTING_BB   1     98.92 12719 823.93
## - TEAM_PITCHING_BB  1    101.48 12722 823.97
## - TEAM_BASERUN_SB   1    109.27 12730 824.09
## <none>                         12620 824.44
## - TEAM_BATTING_3B   1    147.01 12767 824.65
## - TEAM_BATTING_HBP  1    204.39 12825 825.51
## - TEAM_FIELDING_DP  1    649.12 13269 832.02
## - TEAM_PITCHING_HR  1    812.92 13433 834.36
## - TEAM_PITCHING_SO  1   1262.90 13883 840.66
## - TEAM_FIELDING_E   1   1379.34 14000 842.25
##
## Step:  AIC=823.71
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_3B + TEAM_BATTING_BB +
##     TEAM_BASERUN_SB + TEAM_BATTING_HBP + TEAM_PITCHING_HR + TEAM_PITCHING_BB +
##     TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq   RSS    AIC
## - TEAM_BATTING_BB   1     32.85 12738 822.21
## - TEAM_PITCHING_BB  1     43.42 12748 822.37
## - TEAM_BASERUN_SB   1    105.16 12810 823.29
## <none>                         12705 823.71
## - TEAM_BATTING_3B   1    153.13 12858 824.00
## - TEAM_BATTING_HBP  1    183.82 12888 824.46
## - TEAM_BATTING_H    1    504.11 13209 829.15
## - TEAM_FIELDING_DP  1    602.80 13308 830.57
## - TEAM_PITCHING_HR  1    850.25 13555 834.09
## - TEAM_PITCHING_SO  1   1259.72 13964 839.77
## - TEAM_FIELDING_E   1   1419.39 14124 841.94
##
## Step:  AIC=822.21
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_3B + TEAM_BASERUN_SB +
##     TEAM_BATTING_HBP + TEAM_PITCHING_HR + TEAM_PITCHING_BB +
##     TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq   RSS    AIC
## - TEAM_BASERUN_SB   1    109.99 12848 821.85
## <none>                         12738 822.21
## - TEAM_BATTING_3B   1    156.45 12894 822.54
## - TEAM_BATTING_HBP  1    186.58 12924 822.98
## - TEAM_BATTING_H    1    485.67 13223 827.35
## - TEAM_FIELDING_DP  1    623.19 13361 829.33
## - TEAM_PITCHING_HR  1    843.83 13581 832.46
```

```
## - TEAM_PITCHING_SO  1   1267.25 14005 838.32
## - TEAM_FIELDING_E   1   1395.02 14133 840.06
## - TEAM_PITCHING_BB  1   2364.81 15102 852.73
##
## Step:  AIC=821.85
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_3B + TEAM_BATTING_HBP +
##     TEAM_PITCHING_HR + TEAM_PITCHING_BB + TEAM_PITCHING_SO +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                    Df Sum of Sq   RSS    AIC
## - TEAM_BATTING_3B   1    133.47 12981 821.82
## <none>                          12848 821.85
## - TEAM_BATTING_HBP  1    177.11 13025 822.46
## - TEAM_BATTING_H    1    566.11 13414 828.09
## - TEAM_FIELDING_DP  1    737.46 13585 830.51
## - TEAM_PITCHING_HR  1    756.49 13604 830.78
## - TEAM_PITCHING_SO  1   1257.91 14106 837.69
## - TEAM_FIELDING_E   1   1330.40 14178 838.67
## - TEAM_PITCHING_BB  1   2371.12 15219 852.20
##
## Step:  AIC=821.82
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_HBP + TEAM_PITCHING_HR +
##     TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                    Df Sum of Sq   RSS    AIC
## <none>                          12981 821.82
## - TEAM_BATTING_HBP  1    228.70 13210 823.16
## - TEAM_BATTING_H    1    449.87 13431 826.33
## - TEAM_FIELDING_DP  1    813.17 13794 831.43
## - TEAM_PITCHING_HR  1    990.20 13971 833.86
## - TEAM_PITCHING_SO  1   1316.56 14298 838.27
## - TEAM_FIELDING_E   1   1334.60 14316 838.52
## - TEAM_PITCHING_BB  1   2583.00 15564 854.49
```

```
summary(stepFull)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_HBP +
##     TEAM_PITCHING_HR + TEAM_PITCHING_BB + TEAM_PITCHING_SO +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP, data = trainingDataRaw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.2248  -5.6294  -0.0212   5.0439  21.3065
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      60.95454   19.10292   3.191 0.001670 **
## TEAM_BATTING_H    0.02541    0.01009   2.518 0.012648 *
## TEAM_BATTING_HBP  0.08712    0.04852   1.796 0.074211 .
## TEAM_PITCHING_HR  0.08945    0.02394   3.736 0.000249 ***
## TEAM_PITCHING_BB  0.05672    0.00940   6.034 8.66e-09 ***
## TEAM_PITCHING_SO -0.03136    0.00728  -4.308 2.68e-05 ***
## TEAM_FIELDING_E  -0.17218    0.03970  -4.338 2.38e-05 ***
## TEAM_FIELDING_DP -0.11904    0.03516  -3.386 0.000869 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.422 on 183 degrees of freedom
```

```
##   (2085 observations deleted due to missingness)
## Multiple R-squared:  0.5345, Adjusted R-squared:  0.5167
## F-statistic: 30.02 on 7 and 183 DF,  p-value: < 2.2e-16
```

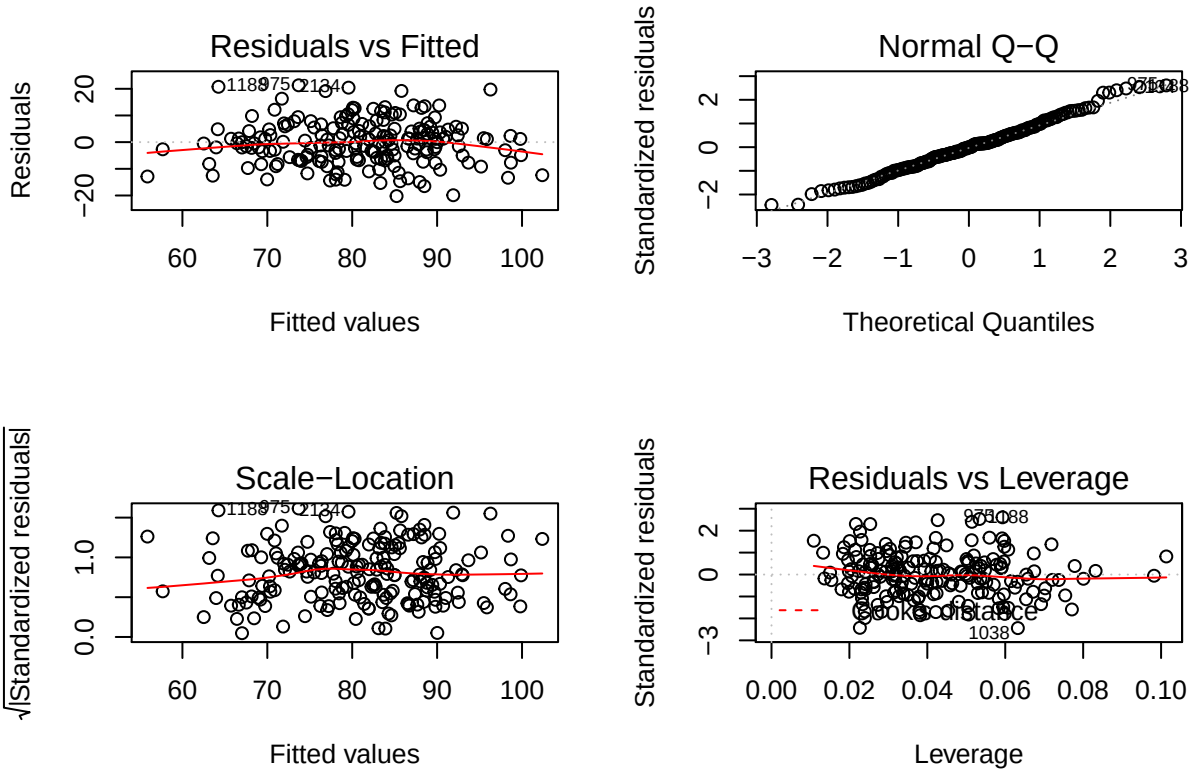
```
#####Generate predictions using the stepFull model
predictionsStepFull <- predict(stepFull, trainingDataRaw)
#View(predictionsStepFull)
```

Generate the RMSE of the stepFull model

```
rmseStep <- sqrt(mean((trainingDataRaw$TARGET_WINS[!is.na(predictionsStepFull)] - predictionsStepFull[!is.na(p
rmseStep
```

```
## [1] 8.244004
```

```
par(mfrow=c(2,2)) #Set up a four panel plot for evaluating regression
plot(stepFull) #Displays Residuals vs Fitted, Scale-Location,  and Normal Q-Q.
```



Evaluation of Stepwise model without TEAM_BATTING_HBP

```
ReducedModel <- lm(TARGET_WINS ~., trainingDataRaw[,c(2:10, 12:17)])
summary(ReducedModel)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = trainingDataRaw[, c(2:10,
##     12:17)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.5627  -6.6932  -0.1328   6.5249  27.8525
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      57.912438   6.642839   8.718  < 2e-16 ***
## TEAM_BATTING_H    0.015434   0.019626   0.786   0.4318
## TEAM_BATTING_2B  -0.070472   0.009369  -7.522 9.36e-14 ***
## TEAM_BATTING_3B   0.161551   0.022192   7.280 5.43e-13 ***
## TEAM_BATTING_HR   0.073952   0.085392   0.866   0.3866
## TEAM_BATTING_BB   0.043765   0.046454   0.942   0.3463
## TEAM_BATTING_SO   0.018250   0.023463   0.778   0.4368
## TEAM_BASERUN_SB   0.035880   0.008687   4.130 3.83e-05 ***
## TEAM_BASERUN_CS   0.052124   0.018227   2.860   0.0043 **
## TEAM_PITCHING_H   0.019044   0.018381   1.036   0.3003
## TEAM_PITCHING_HR  0.022997   0.082092   0.280   0.7794
## TEAM_PITCHING_BB -0.004180   0.044692  -0.094   0.9255
## TEAM_PITCHING_SO -0.038176   0.022447  -1.701   0.0892 .
## TEAM_FIELDING_E  -0.155876   0.009946 -15.672  < 2e-16 ***
## TEAM_FIELDING_DP -0.112885   0.013137  -8.593  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.556 on 1471 degrees of freedom
##   (790 observations deleted due to missingness)
## Multiple R-squared:  0.4386, Adjusted R-squared:  0.4333
## F-statistic:  82.1 on 14 and 1471 DF,  p-value: < 2.2e-16
```

```
stepReduced <- step(ReducedModel)
```

```
## Start:  AIC=6723.18
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_BASERUN_CS + TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_BB +
##     TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq    RSS    AIC
## - TEAM_PITCHING_BB  1       0.8 134324 6721.2
## - TEAM_PITCHING_HR  1       7.2 134330 6721.3
## - TEAM_BATTING_SO   1      55.2 134378 6721.8
## - TEAM_BATTING_H    1      56.5 134380 6721.8
## - TEAM_BATTING_HR   1      68.5 134392 6721.9
## - TEAM_BATTING_BB   1      81.0 134404 6722.1
## - TEAM_PITCHING_H   1      98.0 134421 6722.3
## <none>                          134323 6723.2
## - TEAM_PITCHING_SO  1     264.1 134587 6724.1
## - TEAM_BASERUN_CS   1     746.8 135070 6729.4
## - TEAM_BASERUN_SB   1    1557.8 135881 6738.3
## - TEAM_BATTING_3B   1    4838.9 139162 6773.8
## - TEAM_BATTING_2B   1    5166.3 139489 6777.3
## - TEAM_FIELDING_DP  1    6742.5 141066 6794.0
## - TEAM_FIELDING_E   1   22427.4 156751 6950.6
##
## Step:  AIC=6721.19
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_BASERUN_CS + TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_SO +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq    RSS    AIC
## - TEAM_PITCHING_HR  1       6.4 134330 6719.3
## - TEAM_BATTING_SO   1      56.2 134380 6719.8
```

```
## - TEAM_BATTING_HR   1      77.9 134402 6720.1
## - TEAM_BATTING_H    1     147.2 134471 6720.8
## <none>                          134324 6721.2
## - TEAM_PITCHING_H   1     197.5 134521 6721.4
## - TEAM_PITCHING_SO  1     266.3 134590 6722.1
## - TEAM_BASERUN_CS   1     746.5 135070 6727.4
## - TEAM_BASERUN_SB   1    1564.2 135888 6736.4
## - TEAM_BATTING_3B   1    4840.8 139165 6771.8
## - TEAM_BATTING_2B   1    5175.9 139500 6775.4
## - TEAM_FIELDING_DP  1    6744.6 141069 6792.0
## - TEAM_BATTING_BB   1   12568.9 146893 6852.1
## - TEAM_FIELDING_E   1   22491.7 156816 6949.2
##
## Step:  AIC=6719.26
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_BASERUN_CS + TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP
##
##                    Df Sum of Sq    RSS    AIC
## - TEAM_BATTING_SO   1      51.2 134382 6717.8
## - TEAM_BATTING_H    1     144.7 134475 6718.9
## <none>                          134330 6719.3
## - TEAM_PITCHING_H   1     202.0 134532 6719.5
## - TEAM_PITCHING_SO  1     298.0 134628 6720.6
## - TEAM_BASERUN_CS   1     742.6 135073 6725.5
## - TEAM_BASERUN_SB   1    1570.4 135901 6734.5
## - TEAM_BATTING_3B   1    4842.6 139173 6769.9
## - TEAM_BATTING_2B   1    5198.7 139529 6773.7
## - TEAM_FIELDING_DP  1    6744.4 141075 6790.1
## - TEAM_BATTING_HR   1    9780.8 144111 6821.7
## - TEAM_BATTING_BB   1   12606.9 146937 6850.6
## - TEAM_FIELDING_E   1   22525.1 156855 6947.6
##
## Step:  AIC=6717.83
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BASERUN_SB + TEAM_BASERUN_CS +
##     TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                    Df Sum of Sq    RSS    AIC
## <none>                          134382 6717.8
## - TEAM_BASERUN_CS   1     737.6 135119 6724.0
## - TEAM_PITCHING_H   1    1355.1 135737 6730.7
## - TEAM_BASERUN_SB   1    1575.6 135957 6733.2
## - TEAM_BATTING_H    1    1740.1 136122 6734.9
## - TEAM_BATTING_3B   1    4849.8 139231 6768.5
## - TEAM_BATTING_2B   1    5148.1 139530 6771.7
## - TEAM_FIELDING_DP  1    6779.2 141161 6789.0
## - TEAM_PITCHING_SO  1    7395.1 141777 6795.4
## - TEAM_BATTING_HR   1    9785.1 144167 6820.3
## - TEAM_BATTING_BB   1   12619.7 147001 6849.2
## - TEAM_FIELDING_E   1   22552.0 156934 6946.4
```

```
predictionsStepReduced <- predict(stepReduced, trainingDataRaw[,c(2:10, 12:17)])
rmseStepR <- sqrt(mean((trainingDataRaw$TARGET_WINS[!is.na(predictionsStepReduced)] - predictionsStepReduced[!
rmseStepR
```

```
## [1] 9.509561
```

Aadi Models

```
#all variables that have positive impact
modeldata_positive = trainingdata_bk[, c(2, 3:7, 9, 15, 17)]  #HBP not included
model3 = lm(TARGET_WINS ~., data = modeldata_positive)
summary(model3)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = modeldata_positive)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.593  -7.317   0.373   7.441  33.078
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      24.265684   5.702831   4.255 2.20e-05 ***
## TEAM_BATTING_H    0.028228   0.004436   6.364 2.48e-10 ***
## TEAM_BATTING_2B  -0.031065   0.009280  -3.347 0.000832 ***
## TEAM_BATTING_3B   0.105593   0.019550   5.401 7.48e-08 ***
## TEAM_BATTING_HR   0.119457   0.008952  13.344  < 2e-16 ***
## TEAM_BATTING_BB   0.038976   0.003328  11.711  < 2e-16 ***
## TEAM_BASERUN_SB   0.055463   0.005774   9.606  < 2e-16 ***
## TEAM_PITCHING_SO -0.013314   0.002044  -6.513 9.48e-11 ***
## TEAM_FIELDING_DP -0.073396   0.012858  -5.708 1.33e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.93 on 1826 degrees of freedom
##   (441 observations deleted due to missingness)
## Multiple R-squared:  0.3131, Adjusted R-squared:  0.3101
## F-statistic:   104 on 8 and 1826 DF,  p-value: < 2.2e-16
```

```
#all variables that have negative impact
modeldata_negative = trainingdata_bk[, c(2, 8, 10, 12:14, 16)]
model4 = lm(TARGET_WINS ~., data = modeldata_negative)
summary(model4)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = modeldata_negative)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.958  -7.633   0.486   7.964  63.621
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      76.2047723  2.8076163  27.142  < 2e-16 ***
## TEAM_BATTING_SO  -0.0266353  0.0020304 -13.118  < 2e-16 ***
## TEAM_BASERUN_CS   0.0984814  0.0145750   6.757 2.01e-11 ***
## TEAM_PITCHING_H  -0.0001897  0.0004563  -0.416    0.678
## TEAM_PITCHING_HR  0.1325371  0.0086219  15.372  < 2e-16 ***
## TEAM_PITCHING_BB  0.0146080  0.0032827   4.450 9.22e-06 ***
## TEAM_FIELDING_E  -0.0248176  0.0044858  -5.533 3.72e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 11.66 on 1497 degrees of freedom
##   (772 observations deleted due to missingness)
## Multiple R-squared:  0.2527, Adjusted R-squared:  0.2497
## F-statistic: 84.35 on 6 and 1497 DF,  p-value: < 2.2e-16
```

```
#all batting variables
modeldata_batting = trainingdata_bk[, c(2, 3:7, 8)] #HBP not included
model5 = lm(TARGET_WINS ~., data = modeldata_batting)
summary(model5)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = modeldata_batting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.267  -8.530   0.546   8.894  57.046
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -5.517212   5.034107  -1.096    0.273
## TEAM_BATTING_H   0.042970   0.003702  11.607  < 2e-16 ***
## TEAM_BATTING_2B -0.014056   0.009340  -1.505    0.132
## TEAM_BATTING_3B  0.094165   0.016326   5.768 9.18e-09 ***
## TEAM_BATTING_HR  0.051493   0.009647   5.338 1.04e-07 ***
## TEAM_BATTING_BB  0.027042   0.002782   9.721  < 2e-16 ***
## TEAM_BATTING_SO  0.003040   0.002244   1.354    0.176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.48 on 2167 degrees of freedom
##   (102 observations deleted due to missingness)
## Multiple R-squared:  0.2534, Adjusted R-squared:  0.2513
## F-statistic: 122.6 on 6 and 2167 DF,  p-value: < 2.2e-16
```

```
#all non batting variables
modeldata_notbatting = trainingdata_bk[, c(2, 9:10, 12:17)]
model6 = lm(TARGET_WINS ~., data = modeldata_notbatting)
summary(model6)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = modeldata_notbatting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.861  -7.183   0.136   6.636  32.266
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      103.930980   3.885553  26.748  < 2e-16 ***
## TEAM_BASERUN_SB    0.046815   0.008874   5.276 1.52e-07 ***
## TEAM_BASERUN_CS    0.053835   0.018940   2.842  0.00454 **
## TEAM_PITCHING_H    0.004869   0.001846   2.637  0.00845 **
## TEAM_PITCHING_HR   0.112016   0.008288  13.515  < 2e-16 ***
## TEAM_PITCHING_BB   0.028093   0.003219   8.728  < 2e-16 ***
## TEAM_PITCHING_SO  -0.038118   0.001828 -20.847  < 2e-16 ***
## TEAM_FIELDING_E   -0.130395   0.009931 -13.131  < 2e-16 ***
```

```
## TEAM_FIELDING_DP  -0.107946   0.013795  -7.825 9.60e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.07 on 1477 degrees of freedom
##   (790 observations deleted due to missingness)
## Multiple R-squared:  0.3746, Adjusted R-squared:  0.3712
## F-statistic: 110.6 on 8 and 1477 DF,  p-value: < 2.2e-16
```

```
#most highly correlated with target wins
modeldata_hc = trainingdata_bk[,subset(data_exp, data_exp[, 4] > 0.15)$VAR_NAME]
modeldata_hc = cbind(trainingdata_bk$TARGET_WINS, modeldata_hc)
names(modeldata_hc)[1] = "TARGET_WINS"
model7 = lm(TARGET_WINS ~., data = modeldata_hc)
summary(model7)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = modeldata_hc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.130  -8.853   0.384   9.137  51.367
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.858632   3.444038  -0.830   0.4066
## TEAM_BATTING_2B  -0.017802   0.008965  -1.986   0.0472 *
## TEAM_BATTING_BB   0.029733   0.002832  10.499   <2e-16 ***
## TEAM_BATTING_H    0.048384   0.002724  17.764   <2e-16 ***
## TEAM_BATTING_HR   0.047957   0.021926   2.187   0.0288 *
## TEAM_PITCHING_HR -0.026827   0.020759  -1.292   0.1964
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.89 on 2270 degrees of freedom
## Multiple R-squared:  0.224,  Adjusted R-squared:  0.2222
## F-statistic:   131 on 5 and 2270 DF,  p-value: < 2.2e-16
```

```
modeldata_hc = trainingdata_bk[,subset(data_exp, data_exp[, 4] < 0)$VAR_NAME]
modeldata_hc = cbind(trainingdata_bk$TARGET_WINS, modeldata_hc)
names(modeldata_hc)[1] = "TARGET_WINS"
model8 = lm(TARGET_WINS ~., data = modeldata_hc)
summary(model8)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = modeldata_hc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.534  -8.412   0.202   8.503  39.946
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      76.182281   3.314972  22.981  < 2e-16 ***
## TEAM_BATTING_SO   0.028201   0.004936   5.713 1.29e-08 ***
## TEAM_FIELDING_DP -0.113491   0.013209  -8.592  < 2e-16 ***
## TEAM_FIELDING_E  -0.081888   0.004267 -19.190  < 2e-16 ***
```

```
## TEAM_PITCHING_H   0.028134   0.001473  19.096  < 2e-16 ***
## TEAM_PITCHING_SO -0.036648   0.004381  -8.364  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.08 on 1882 degrees of freedom
##   (388 observations deleted due to missingness)
## Multiple R-squared:  0.2137, Adjusted R-squared:  0.2116
## F-statistic: 102.3 on 5 and 1882 DF,  p-value: < 2.2e-16
```

```
completecases = trainingdata_bk[complete.cases(trainingdata_bk),]
model9 = lm(TARGET_WINS ~., data = completecases)
summary(model9)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = completecases)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.0626  -5.4196  -0.0423   5.2111  22.9355
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      60.4562317 19.7385030   3.063  0.00254 **
## INDEX            -0.0002478  0.0008508  -0.291  0.77122
## TEAM_BATTING_H    1.8111103  2.7908648   0.649  0.51723
## TEAM_BATTING_2B   0.0267462  0.0303941   0.880  0.38008
## TEAM_BATTING_3B  -0.1018043  0.0777401  -1.310  0.19208
## TEAM_BATTING_HR  -4.6100155 10.5666083  -0.436  0.66317
## TEAM_BATTING_BB  -4.4606275  3.6457882  -1.224  0.22279
## TEAM_BATTING_SO   0.4303282  2.6231874   0.164  0.86988
## TEAM_BASERUN_SB   0.0335937  0.0288100   1.166  0.24519
## TEAM_BASERUN_CS  -0.0130338  0.0719436  -0.181  0.85645
## TEAM_BATTING_HBP  0.0837038  0.0499097   1.677  0.09532 .
## TEAM_PITCHING_H  -1.7887761  2.7903398  -0.641  0.52233
## TEAM_PITCHING_HR  4.6958245 10.5649821   0.444  0.65725
## TEAM_PITCHING_BB  4.5120283  3.6432611   1.238  0.21721
## TEAM_PITCHING_SO -0.4618971  2.6214432  -0.176  0.86034
## TEAM_FIELDING_E  -0.1724513  0.0415365  -4.152 5.16e-05 ***
## TEAM_FIELDING_DP -0.1063200  0.0371964  -2.858  0.00478 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.489 on 174 degrees of freedom
## Multiple R-squared:  0.5503, Adjusted R-squared:  0.509
## F-statistic: 13.31 on 16 and 174 DF,  p-value: < 2.2e-16
```

Model Selection and Prediction