Homework #1: Baseball Analysis

Data 621 Business Analytics and Data Mining

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

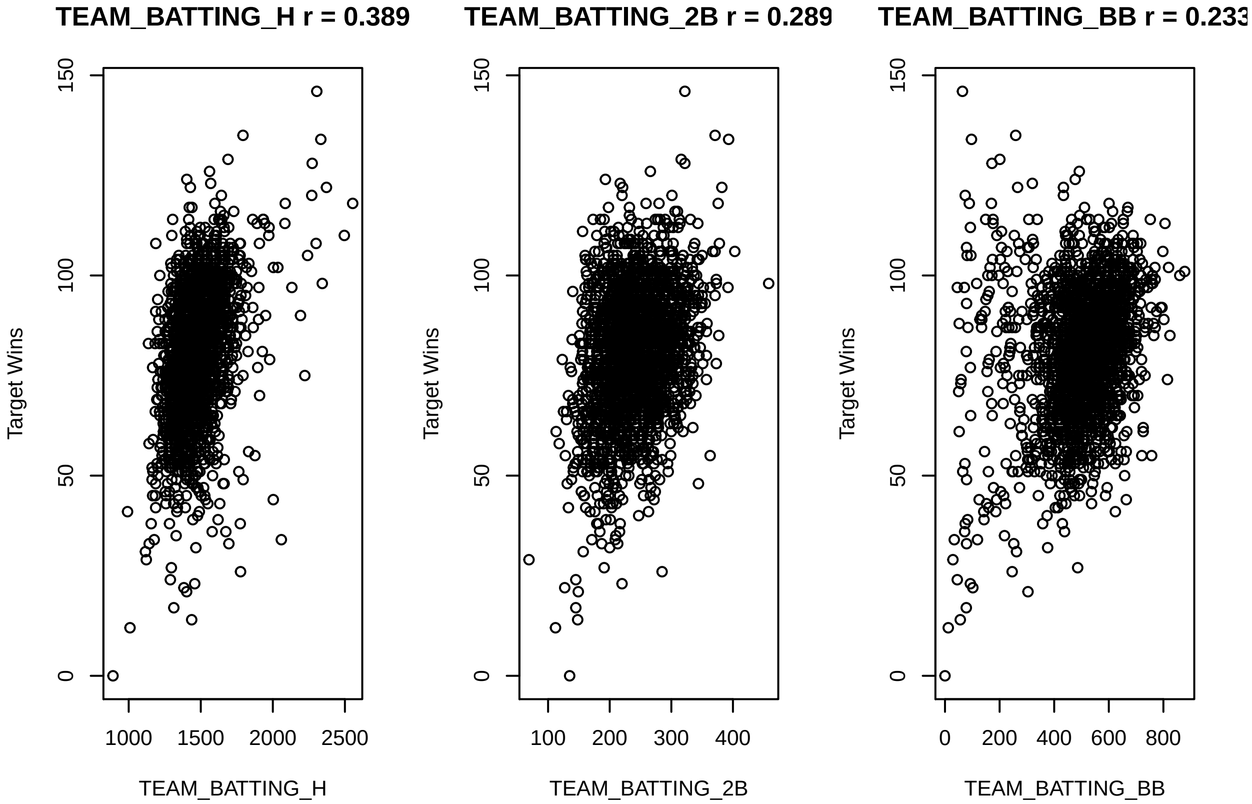
Due June 19, 2016

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	TARGET_WINS	80.79086	82.0	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_H	1469.26977	1454.0	0.3887675	0
7	TEAM_BATTING_HBP	58.11380	58.0	0.0165164	0
8	TEAM_BATTING_HR	99.61204	102.0	0.1761532	0
9	TEAM_BATTING_SO	736.25044	750.0	-0.0305814	0
10	TEAM_FIELDING_DP	146.71617	149.0	-0.0300863	0
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	817.54086	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%	7%	31%
b	TEAM_BASERUN_CS	Strikeouts by batters	Negative	33.9%	2%	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

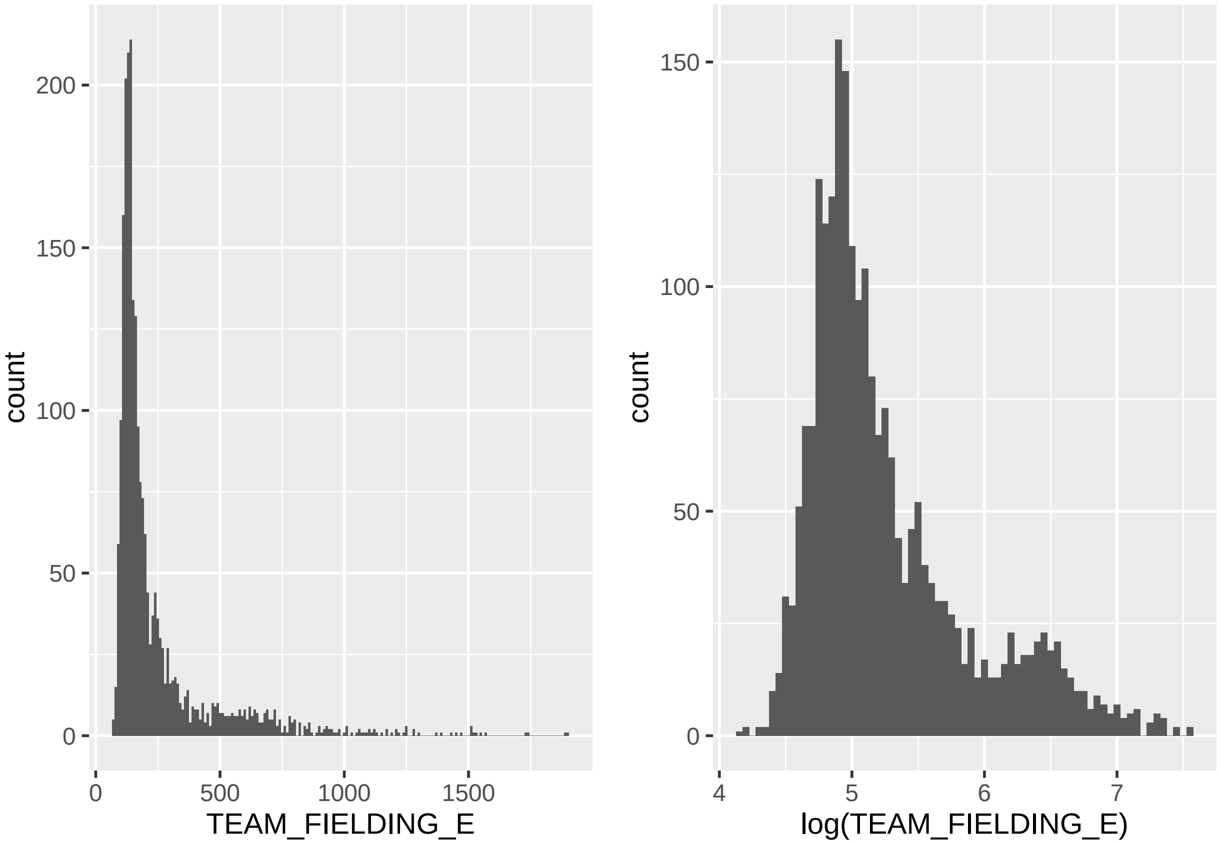
Use all the variables to see p value of each variables.

```
fit_all <- lm(TARGET_WINS ~ . , df_new)
summary(fit_all)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = df_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.821  -8.616   0.068   8.289  59.070
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      21.3843494  6.7992701   3.145  0.00168 **
## INDEX            -0.0004530  0.0003766  -1.203  0.22918
## TEAM_BATTING_H    0.0488087  0.0036959  13.206  < 2e-16 ***
## TEAM_BATTING_2B  -0.0210728  0.0091813  -2.295  0.02181 *
## TEAM_BATTING_3B   0.0656929  0.0168328   3.903 9.79e-05 ***
## TEAM_BATTING_HR   0.0531978  0.0275007   1.934  0.05319 .
## TEAM_BATTING_BB   0.0102316  0.0058407   1.752  0.07995 .
## TEAM_BATTING_SO  -0.0083756  0.0025502  -3.284  0.00104 **
## TEAM_BASERUN_SB   0.0257931  0.0043664   5.907 4.01e-09 ***
## TEAM_BASERUN_CS  -0.0108216  0.0157870  -0.685  0.49312
## TEAM_BATTING_HBP  0.0487185  0.0730953   0.667  0.50516
## TEAM_PITCHING_H  -0.0008239  0.0003678  -2.240  0.02518 *
## TEAM_PITCHING_HR  0.0129919  0.0243930   0.533  0.59436
## TEAM_PITCHING_BB  0.0006724  0.0041580   0.162  0.87154
## TEAM_PITCHING_SO  0.0028321  0.0009221   3.071  0.00216 **
## TEAM_FIELDING_E  -0.0196745  0.0024632  -7.987 2.18e-15 ***
## TEAM_FIELDING_DP -0.1209399  0.0129572  -9.334  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.07 on 2259 degrees of freedom
## Multiple R-squared:  0.3159, Adjusted R-squared:  0.3111
## F-statistic: 65.21 on 16 and 2259 DF,  p-value: < 2.2e-16
```

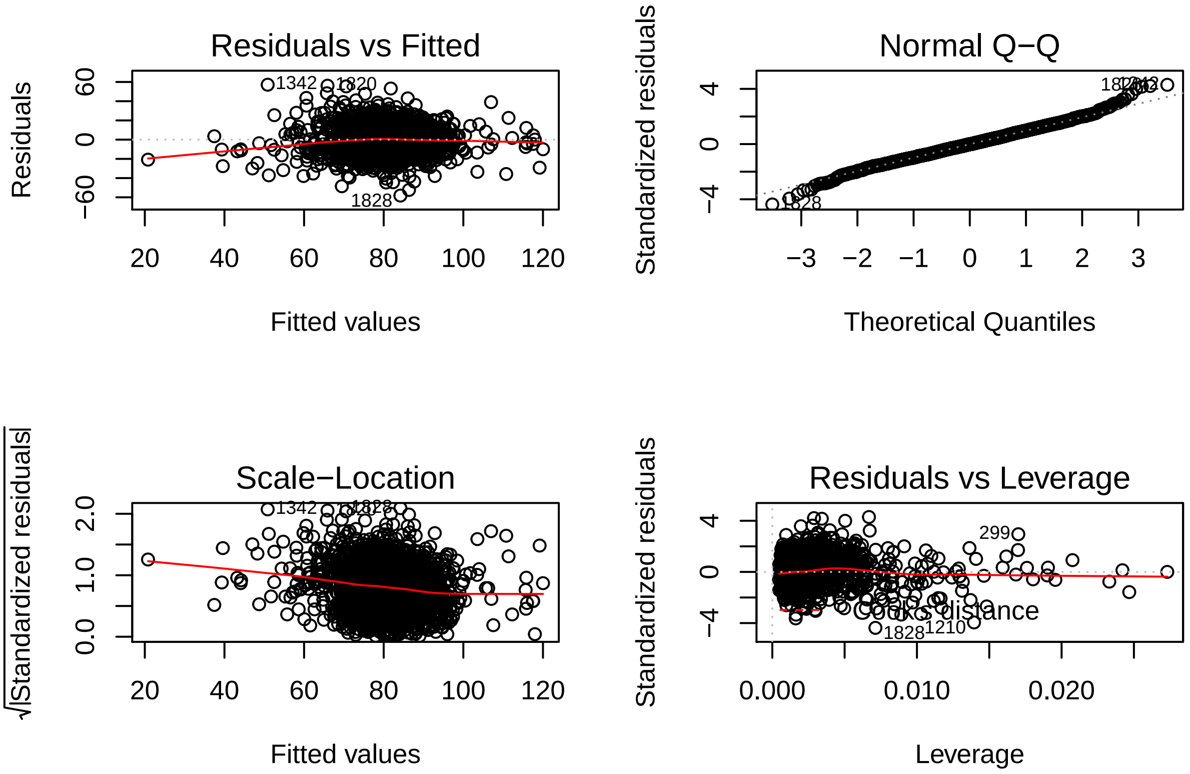
Model –(Nathan)

```
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)
```

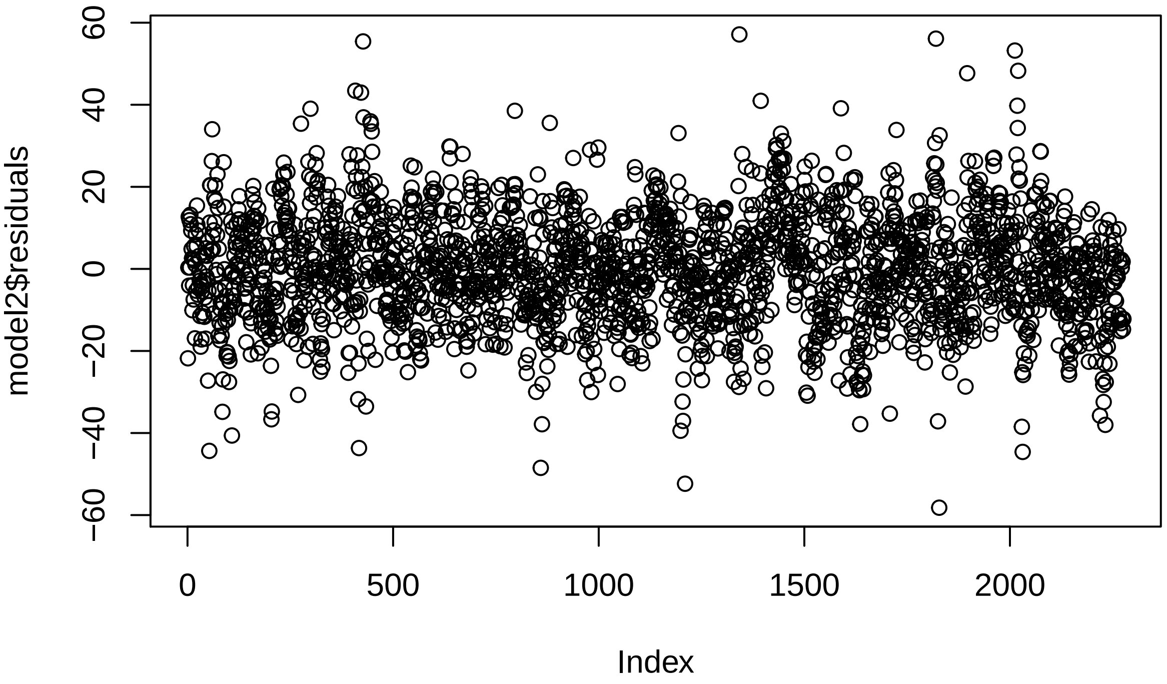


```
model2<- lm(TARGET_WINS ~  TEAM_BATTING_H + TEAM_BASERUN_SB + TEAM_FIELDING_DP +
              log(TEAM_FIELDING_E), df_new)

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



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



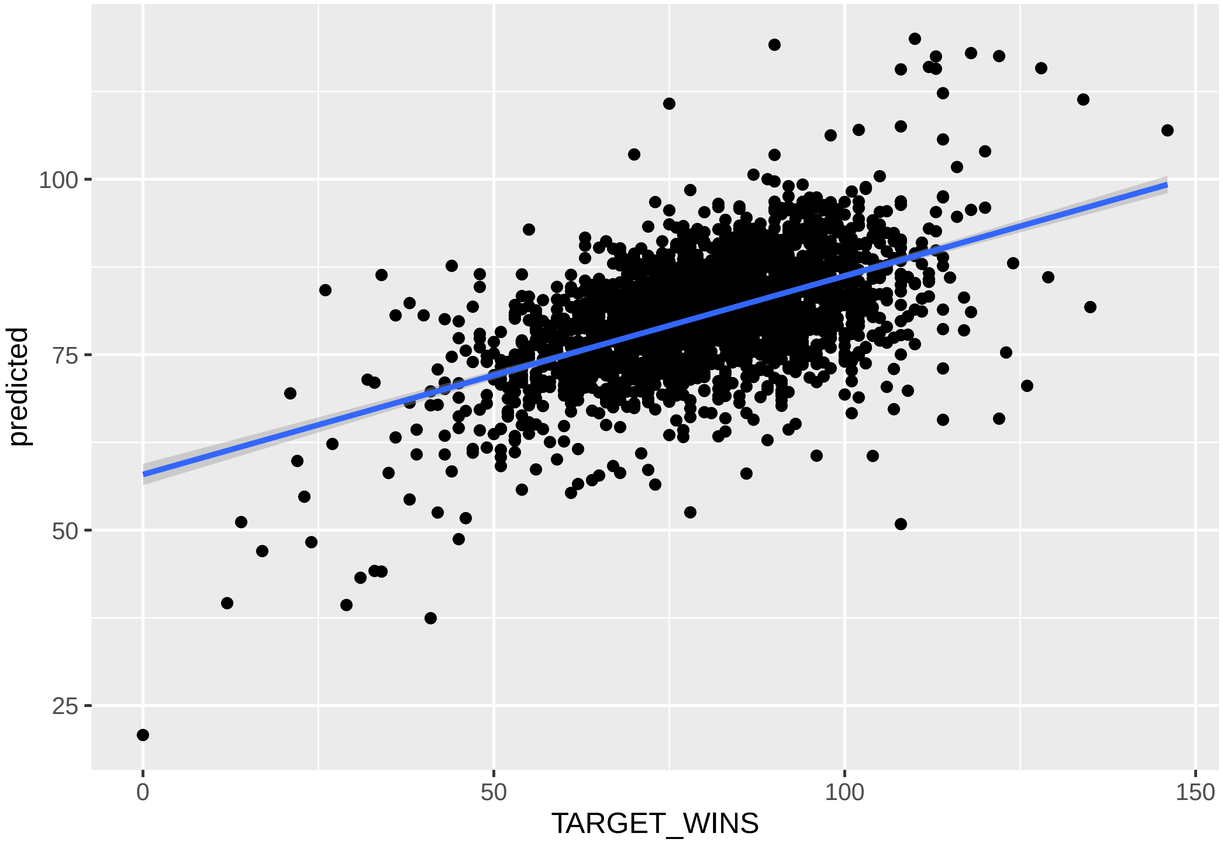
```r
summary(model2)
```

```
##
## 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
```

```
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"
```

```
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]

```
#dfraw <- read.csv(url("https://raw.githubusercontent.com/dsmilo/DATA621/master/HW1/data/moneyball-training-da
dfraw = trainingdata
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
## -45.940  -8.119   0.083   7.991  69.074
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      30.337651   5.861651   5.176 2.48e-07 ***
## TEAM_BATTING_H    0.007809   0.005342   1.462 0.143892
## TEAM_BATTING_2B  -0.015508   0.009208  -1.684 0.092281 .
## TEAM_BATTING_3B   0.156682   0.018619   8.415  < 2e-16 ***
## TEAM_BATTING_HR   0.080380   0.054071   1.487 0.137272
## TEAM_BATTING_BB   0.080952   0.016755   4.831 1.45e-06 ***
## TEAM_BATTING_SO  -0.004743   0.002360  -2.009 0.044619 *
## TEAM_BASERUN_SB   0.038907   0.004484   8.677  < 2e-16 ***
## TEAM_PITCHING_H   0.022266   0.002364   9.420  < 2e-16 ***
## TEAM_PITCHING_HR -0.007882   0.050655  -0.156 0.876358
## TEAM_PITCHING_BB -0.054024   0.015004  -3.601 0.000324 ***
## TEAM_FIELDING_E  -0.038008   0.003552 -10.700  < 2e-16 ***
## TEAM_FIELDING_DP -0.089224   0.012726  -7.011 3.14e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.39 on 2186 degrees of freedom
## Multiple R-squared:  0.2867, Adjusted R-squared:  0.2828
## F-statistic: 73.24 on 12 and 2186 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
## -47.276  -8.108   0.155   7.931  70.414
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      36.157377   4.303766   8.401  < 2e-16 ***
## TEAM_BATTING_2B  -0.008123   0.007701  -1.055  0.29159    
## TEAM_BATTING_3B   0.167156   0.017190   9.724  < 2e-16 ***
## TEAM_BATTING_HR   0.078505   0.054069   1.452  0.14666    
## TEAM_BATTING_BB   0.089644   0.015669   5.721 1.20e-08 ***
## TEAM_BATTING_SO  -0.006312   0.002103  -3.002  0.00271 ** 
## TEAM_BASERUN_SB   0.041200   0.004202   9.806  < 2e-16 ***
## TEAM_PITCHING_H   0.024375   0.001872  13.018  < 2e-16 ***
## TEAM_PITCHING_HR -0.001457   0.050477  -0.029  0.97697    
## TEAM_PITCHING_BB -0.061865   0.014016  -4.414 1.06e-05 ***
## TEAM_FIELDING_E  -0.039317   0.003438 -11.435  < 2e-16 ***
## TEAM_FIELDING_DP -0.086472   0.012590  -6.869 8.42e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 12.39 on 2187 degrees of freedom
## Multiple R-squared:  0.286,  Adjusted R-squared:  0.2825 
## F-statistic: 79.66 on 11 and 2187 DF,  p-value: < 2.2e-16
```

```
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 
## -47.248  -8.106   0.156   7.929  70.391 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      36.140217   4.261553   8.481  < 2e-16 ***
## TEAM_BATTING_2B  -0.008135   0.007688  -1.058   0.2901    
## TEAM_BATTING_3B   0.167090   0.017033   9.810  < 2e-16 ***
## TEAM_BATTING_HR   0.076966   0.008975   8.576  < 2e-16 ***
## TEAM_BATTING_BB   0.089990   0.010109   8.902  < 2e-16 ***
## TEAM_BATTING_SO  -0.006310   0.002101  -3.003   0.0027 ** 
## TEAM_BASERUN_SB   0.041207   0.004194   9.825  < 2e-16 ***
## TEAM_PITCHING_H   0.024386   0.001837  13.272  < 2e-16 ***
## TEAM_PITCHING_BB -0.062186   0.008498  -7.318 3.53e-13 ***
## TEAM_FIELDING_E  -0.039286   0.003264 -12.038  < 2e-16 ***
## TEAM_FIELDING_DP -0.086480   0.012583  -6.873 8.19e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 12.39 on 2188 degrees of freedom
## Multiple R-squared:  0.286,  Adjusted R-squared:  0.2828 
## F-statistic: 87.66 on 10 and 2188 DF,  p-value: < 2.2e-16
```

```
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
## -46.531  -8.109   0.129   7.876  69.256
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      36.263163   4.260085   8.512  < 2e-16 ***
## TEAM_BATTING_3B   0.164696   0.016882   9.755  < 2e-16 ***
## TEAM_BATTING_HR   0.075346   0.008844   8.520  < 2e-16 ***
## TEAM_BATTING_BB   0.086862   0.009667   8.985  < 2e-16 ***
## TEAM_BATTING_SO  -0.006435   0.002098  -3.068  0.00218 **
## TEAM_BASERUN_SB   0.040964   0.004188   9.782  < 2e-16 ***
## TEAM_PITCHING_H   0.023468   0.001620  14.486  < 2e-16 ***
## TEAM_PITCHING_BB -0.059647   0.008153  -7.316 3.56e-13 ***
## TEAM_FIELDING_E  -0.038321   0.003133 -12.229  < 2e-16 ***
## TEAM_FIELDING_DP -0.087880   0.012514  -7.023 2.89e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.39 on 2189 degrees of freedom
## Multiple R-squared:  0.2857, Adjusted R-squared:  0.2827
## F-statistic: 97.27 on 9 and 2189 DF,  p-value: < 2.2e-16
```

```r
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
## -45.969  -8.277   0.246   8.056  73.242
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      51.559032   3.756061  13.727  < 2e-16 ***
## TEAM_BATTING_3B   0.176739   0.017002  10.395  < 2e-16 ***
## TEAM_BATTING_HR   0.092323   0.008636  10.691  < 2e-16 ***
## TEAM_BATTING_BB   0.020446   0.003363   6.079 1.42e-09 ***
## TEAM_BATTING_SO  -0.008490   0.002104  -4.036 5.62e-05 ***
## TEAM_BASERUN_SB   0.036911   0.004200   8.787  < 2e-16 ***
## TEAM_PITCHING_H   0.015277   0.001185  12.894  < 2e-16 ***
## TEAM_FIELDING_E  -0.040639   0.003155 -12.883  < 2e-16 ***
## TEAM_FIELDING_DP -0.093260   0.012641  -7.377 2.28e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.54 on 2190 degrees of freedom
## Multiple R-squared:  0.2682, Adjusted R-squared:  0.2655
## F-statistic: 100.3 on 8 and 2190 DF,  p-value: < 2.2e-16
```

```r
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
## -46.040  -8.414   0.287   8.311  75.045
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      78.057639   3.260337  23.942  < 2e-16 ***
## TEAM_BATTING_3B   0.190274   0.017598  10.812  < 2e-16 ***
## TEAM_BATTING_HR   0.133955   0.008306  16.128  < 2e-16 ***
## TEAM_BATTING_BB   0.014287   0.003453   4.138 3.64e-05 ***
## TEAM_BATTING_SO  -0.016697   0.002079  -8.030 1.57e-15 ***
## TEAM_BASERUN_SB   0.028286   0.004300   6.577 5.97e-11 ***
## TEAM_FIELDING_E  -0.016658   0.002642  -6.304 3.49e-10 ***
## TEAM_FIELDING_DP -0.106359   0.013067  -8.140 6.58e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13 on 2191 degrees of freedom
## Multiple R-squared:  0.2127, Adjusted R-squared:  0.2101
## F-statistic: 84.54 on 7 and 2191 DF,  p-value: < 2.2e-16
```

```r
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
## -46.165  -8.605   0.306   8.585  72.867
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     63.900645   2.798411  22.835  < 2e-16 ***
## TEAM_BATTING_3B  0.195985   0.017844  10.983  < 2e-16 ***
## TEAM_BATTING_HR  0.116108   0.008129  14.282  < 2e-16 ***
## TEAM_BATTING_BB  0.008770   0.003435   2.553   0.0108 *
## TEAM_BATTING_SO -0.013440   0.002071  -6.491 1.05e-10 ***
## TEAM_BASERUN_SB  0.033092   0.004323   7.656 2.87e-14 ***
## TEAM_FIELDING_E -0.017434   0.002680  -6.506 9.51e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.19 on 2192 degrees of freedom
## Multiple R-squared:  0.1889, Adjusted R-squared:  0.1866
## F-statistic: 85.06 on 6 and 2192 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
## -49.745  -8.623   0.137   8.390  58.605
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      21.0038417  6.7925780   3.092 0.002011 **
## TEAM_BATTING_H    0.0489011  0.0036954  13.233  < 2e-16 ***
## TEAM_BATTING_2B  -0.0210986  0.0091822  -2.298 0.021666 *
## TEAM_BATTING_3B   0.0645246  0.0168064   3.839 0.000127 ***
## TEAM_BATTING_HR   0.0525039  0.0274974   1.909 0.056335 .
## TEAM_BATTING_BB   0.0104483  0.0058384   1.790 0.073657 .
## TEAM_BATTING_SO  -0.0084975  0.0025484  -3.334 0.000869 ***
## TEAM_BASERUN_SB   0.0254442  0.0043572   5.840 5.99e-09 ***
## TEAM_BASERUN_CS  -0.0108293  0.0157886  -0.686 0.492852
## TEAM_BATTING_HBP  0.0466590  0.0730825   0.638 0.523250
## TEAM_PITCHING_H  -0.0008451  0.0003674  -2.300 0.021540 *
## TEAM_PITCHING_HR  0.0131780  0.0243950   0.540 0.589116
## TEAM_PITCHING_BB  0.0007612  0.0041578   0.183 0.854747
## TEAM_PITCHING_SO  0.0028222  0.0009221   3.061 0.002235 **
## TEAM_FIELDING_E  -0.0195730  0.0024620  -7.950 2.92e-15 ***
## TEAM_FIELDING_DP -0.1215789  0.0129476  -9.390  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.08 on 2260 degrees of freedom
## Multiple R-squared:  0.3155, Adjusted R-squared:  0.311
## F-statistic: 69.45 on 15 and 2260 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=11717.97
## 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_PITCHING_BB  1       5.7 386399 11716
## - TEAM_PITCHING_HR  1      49.9 386443 11716
## - TEAM_BATTING_HBP  1      69.7 386463 11716
## - TEAM_BASERUN_CS   1      80.4 386473 11716
## <none>                          386393 11718
## - TEAM_BATTING_BB   1     547.5 386940 11719
## - TEAM_BATTING_HR   1     623.3 387016 11720
## - TEAM_BATTING_2B   1     902.7 387295 11721
## - TEAM_PITCHING_H   1     904.4 387297 11721
## - TEAM_PITCHING_SO  1    1601.5 387994 11725
## - TEAM_BATTING_SO   1    1900.9 388294 11727
## - TEAM_BATTING_3B   1    2520.1 388913 11731
## - TEAM_BASERUN_SB   1    5830.3 392223 11750
## - TEAM_FIELDING_E   1   10805.7 397199 11779
## - TEAM_FIELDING_DP  1   15075.0 401468 11803
## - TEAM_BATTING_H    1   29938.2 416331 11886
##
## Step:  AIC=11716.01
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_BASERUN_CS + TEAM_BATTING_HBP + TEAM_PITCHING_H + TEAM_PITCHING_HR +
##     TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                     Df Sum of Sq    RSS   AIC
## - TEAM_BATTING_HBP  1      70.0 386469 11714
## - TEAM_BASERUN_CS   1      82.1 386481 11714
## - TEAM_PITCHING_HR  1      91.9 386490 11714
## <none>                          386399 11716
## - TEAM_BATTING_HR   1     737.3 387136 11718
## - TEAM_BATTING_2B   1     900.9 387299 11719
## - TEAM_PITCHING_H   1    1052.7 387451 11720
## - TEAM_BATTING_BB   1    1903.3 388302 11725
## - TEAM_BATTING_SO   1    2105.2 388504 11726
## - TEAM_BATTING_3B   1    2519.5 388918 11729
## - TEAM_PITCHING_SO  1    3255.1 389654 11733
## - TEAM_BASERUN_SB   1    6025.1 392424 11749
## - TEAM_FIELDING_E   1   10801.2 397200 11777
## - TEAM_FIELDING_DP  1   15069.4 401468 11801
## - TEAM_BATTING_H    1   29979.5 416378 11884
##
## Step:  AIC=11714.42
## 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_BASERUN_CS   1      84.8 386553 11713
## - TEAM_PITCHING_HR  1      90.0 386558 11713
## <none>                          386469 11714
## - TEAM_BATTING_HR   1     742.4 387211 11717
## - TEAM_BATTING_2B   1     889.3 387358 11718
## - TEAM_PITCHING_H   1    1052.2 387521 11719
## - TEAM_BATTING_BB   1    1910.6 388379 11724
## - TEAM_BATTING_SO   1    2078.2 388547 11725
## - TEAM_BATTING_3B   1    2516.0 388984 11727
## - TEAM_PITCHING_SO  1    3247.0 389716 11732
## - TEAM_BASERUN_SB   1    6017.0 392486 11748
## - TEAM_FIELDING_E   1   10763.3 397232 11775
## - TEAM_FIELDING_DP  1   15128.1 401597 11800
## - TEAM_BATTING_H    1   29996.7 416465 11883
##
## Step:  AIC=11712.92
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_SO + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP
##
##                    Df Sum of Sq    RSS   AIC
## - TEAM_PITCHING_HR  1      86.4 386640 11711
## <none>                          386553 11713
## - TEAM_BATTING_HR   1     793.8 387347 11716
## - TEAM_BATTING_2B   1     912.6 387466 11716
## - TEAM_PITCHING_H   1    1080.6 387634 11717
## - TEAM_BATTING_BB   1    2005.6 388559 11723
## - TEAM_BATTING_SO   1    2079.5 388633 11723
## - TEAM_BATTING_3B   1    2555.4 389109 11726
## - TEAM_PITCHING_SO  1    3269.0 389822 11730
## - TEAM_BASERUN_SB   1    5983.2 392536 11746
## - TEAM_FIELDING_E   1   10870.9 397424 11774
## - TEAM_FIELDING_DP  1   15186.6 401740 11799
## - TEAM_BATTING_H    1   29953.0 416506 11881
##
## Step:  AIC=11711.43
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                    Df Sum of Sq    RSS   AIC
## <none>                          386640 11711
## - TEAM_BATTING_2B   1     929.4 387569 11715
## - TEAM_PITCHING_H   1    1001.0 387641 11715
## - TEAM_BATTING_BB   1    1999.1 388639 11721
## - TEAM_BATTING_SO   1    2060.9 388701 11722
## - TEAM_BATTING_3B   1    2739.4 389379 11726
## - TEAM_PITCHING_SO  1    3328.3 389968 11729
## - TEAM_BASERUN_SB   1    5986.1 392626 11744
## - TEAM_BATTING_HR   1    8364.1 395004 11758
## - TEAM_FIELDING_E   1   10786.9 397427 11772
## - TEAM_FIELDING_DP  1   15152.3 401792 11797
## - TEAM_BATTING_H    1   30558.9 417199 11883
```

```
summary(stepFull)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO +
##     TEAM_BASERUN_SB + TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP, data = trainingDataRaw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.598  -8.593   0.085   8.445  58.582
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      22.3440443  5.2338369   4.269 2.04e-05 ***
## TEAM_BATTING_H    0.0490922  0.0036699  13.377  < 2e-16 ***
## TEAM_BATTING_2B  -0.0213744  0.0091626  -2.333 0.019746 *
## TEAM_BATTING_3B   0.0665763  0.0166230   4.005 6.40e-05 ***
## TEAM_BATTING_HR   0.0674046  0.0096315   6.998 3.40e-12 ***
## TEAM_BATTING_BB   0.0115464  0.0033748   3.421 0.000634 ***
## TEAM_BATTING_SO  -0.0085211  0.0024529  -3.474 0.000523 ***
## TEAM_BASERUN_SB   0.0249207  0.0042092   5.920 3.70e-09 ***
## TEAM_PITCHING_H  -0.0007770  0.0003209  -2.421 0.015552 *
## TEAM_PITCHING_SO  0.0029662  0.0006719   4.415 1.06e-05 ***
## TEAM_FIELDING_E  -0.0190100  0.0023919  -7.948 2.97e-15 ***
## TEAM_FIELDING_DP -0.1217894  0.0129296  -9.419  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.07 on 2264 degrees of freedom
## Multiple R-squared:  0.3151, Adjusted R-squared:  0.3117
## F-statistic: 94.68 on 11 and 2264 DF,  p-value: < 2.2e-16
```

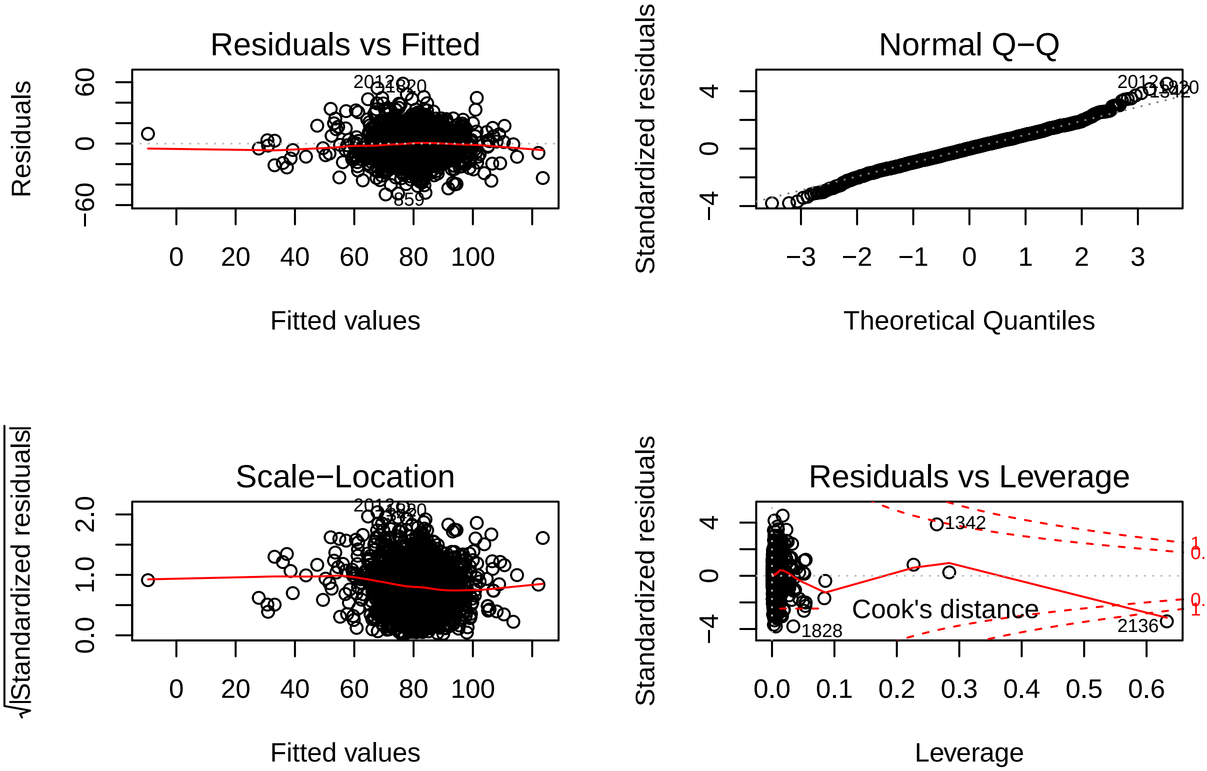
```
#####Generate predictions using the stepFull model
predictionsStepFull <- predict(stepFull, trainingDataRaw)
#View(predictionsStepFull)
```

Generate the RMSE of the stepFull model

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

```
## [1] 13.03368
```

```
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
## -49.753  -8.626   0.120   8.395  58.561
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      23.6421579  5.3902272   4.386 1.21e-05 ***
## TEAM_BATTING_H    0.0489152  0.0036949  13.239  < 2e-16 ***
## TEAM_BATTING_2B  -0.0209575  0.0091783  -2.283 0.022501 *
## TEAM_BATTING_3B   0.0644788  0.0168040   3.837 0.000128 ***
## TEAM_BATTING_HR   0.0527325  0.0274915   1.918 0.055219 .
## TEAM_BATTING_BB   0.0104483  0.0058377   1.790 0.073621 .
## TEAM_BATTING_SO  -0.0084323  0.0025461  -3.312 0.000941 ***
## TEAM_BASERUN_SB   0.0254236  0.0043565   5.836 6.12e-09 ***
## TEAM_BASERUN_CS  -0.0110027  0.0157842  -0.697 0.485829
## TEAM_PITCHING_H  -0.0008456  0.0003674  -2.302 0.021444 *
## TEAM_PITCHING_HR  0.0129626  0.0243894   0.531 0.595135
## TEAM_PITCHING_BB  0.0007798  0.0041571   0.188 0.851231
## TEAM_PITCHING_SO  0.0028156  0.0009219   3.054 0.002284 **
## TEAM_FIELDING_E  -0.0195325  0.0024609  -7.937 3.23e-15 ***
## TEAM_FIELDING_DP -0.1217801  0.0129421  -9.410  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.07 on 2261 degrees of freedom
## Multiple R-squared:  0.3154, Adjusted R-squared:  0.3111
## F-statistic:  74.4 on 14 and 2261 DF,  p-value: < 2.2e-16
```

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

```
## Start:  AIC=11716.38
## 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       6.0 386469 11714
## - TEAM_PITCHING_HR  1      48.3 386511 11715
## - TEAM_BASERUN_CS   1      83.1 386546 11715
## <none>                          386463 11716
## - TEAM_BATTING_BB   1     547.5 387010 11718
## - TEAM_BATTING_HR   1     628.9 387091 11718
## - TEAM_BATTING_2B   1     891.2 387354 11720
## - TEAM_PITCHING_H   1     905.5 387368 11720
## - TEAM_PITCHING_SO  1    1594.2 388057 11724
## - TEAM_BATTING_SO   1    1874.9 388337 11725
## - TEAM_BATTING_3B   1    2516.6 388979 11729
## - TEAM_BASERUN_SB   1    5821.2 392284 11748
## - TEAM_FIELDING_E   1   10768.2 397231 11777
## - TEAM_FIELDING_DP  1   15134.0 401596 11802
## - TEAM_BATTING_H    1   29956.6 416419 11884
##
## Step:  AIC=11714.42
## 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_BASERUN_CS   1      84.8 386553 11713
## - TEAM_PITCHING_HR  1      90.0 386558 11713
## <none>                          386469 11714
## - TEAM_BATTING_HR   1     742.4 387211 11717
## - TEAM_BATTING_2B   1     889.3 387358 11718
## - TEAM_PITCHING_H   1    1052.2 387521 11719
## - TEAM_BATTING_BB   1    1910.6 388379 11724
## - TEAM_BATTING_SO   1    2078.2 388547 11725
## - TEAM_BATTING_3B   1    2516.0 388984 11727
## - TEAM_PITCHING_SO  1    3247.0 389716 11732
## - TEAM_BASERUN_SB   1    6017.0 392486 11748
## - TEAM_FIELDING_E   1   10763.3 397232 11775
## - TEAM_FIELDING_DP  1   15128.1 401597 11800
## - TEAM_BATTING_H    1   29996.7 416465 11883
##
## Step:  AIC=11712.92
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_SO + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP
##
##                    Df Sum of Sq    RSS   AIC
## - TEAM_PITCHING_HR  1      86.4 386640 11711
## <none>                          386553 11713
## - TEAM_BATTING_HR   1     793.8 387347 11716
## - TEAM_BATTING_2B   1     912.6 387466 11716
## - TEAM_PITCHING_H   1    1080.6 387634 11717
## - TEAM_BATTING_BB   1    2005.6 388559 11723
## - TEAM_BATTING_SO   1    2079.5 388633 11723
## - TEAM_BATTING_3B   1    2555.4 389109 11726
## - TEAM_PITCHING_SO  1    3269.0 389822 11730
## - TEAM_BASERUN_SB   1    5983.2 392536 11746
## - TEAM_FIELDING_E   1   10870.9 397424 11774
## - TEAM_FIELDING_DP  1   15186.6 401740 11799
## - TEAM_BATTING_H    1   29953.0 416506 11881
##
## Step:  AIC=11711.43
## TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                    Df Sum of Sq    RSS   AIC
## <none>                          386640 11711
## - TEAM_BATTING_2B   1     929.4 387569 11715
## - TEAM_PITCHING_H   1    1001.0 387641 11715
## - TEAM_BATTING_BB   1    1999.1 388639 11721
## - TEAM_BATTING_SO   1    2060.9 388701 11722
## - TEAM_BATTING_3B   1    2739.4 389379 11726
## - TEAM_PITCHING_SO  1    3328.3 389968 11729
## - TEAM_BASERUN_SB   1    5986.1 392626 11744
## - TEAM_BATTING_HR   1    8364.1 395004 11758
## - TEAM_FIELDING_E   1   10786.9 397427 11772
## - TEAM_FIELDING_DP  1   15152.3 401792 11797
## - TEAM_BATTING_H    1   30558.9 417199 11883
```

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

```
## [1] 13.03368
```

Model Selection and Prediction