Homework #1: Baseball Analysis

Data 621 Business Analytics and Data Mining

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

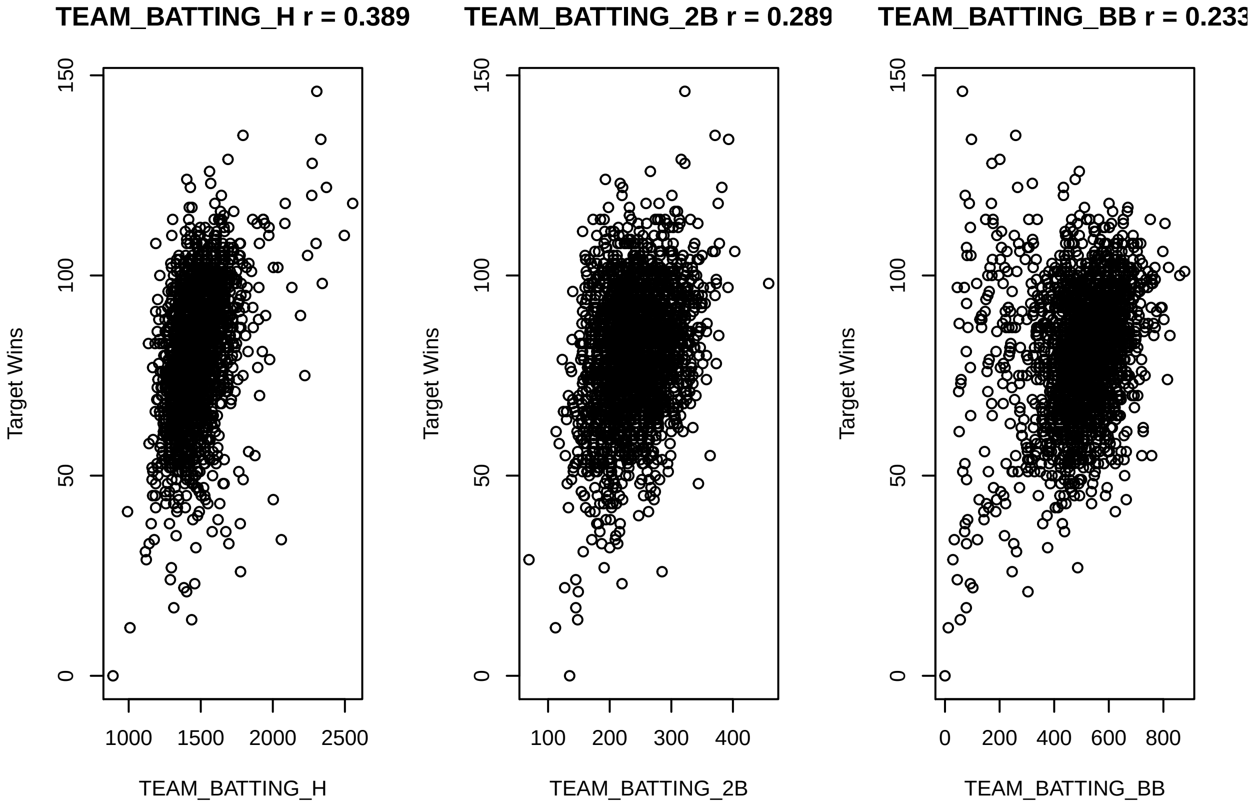
Due June 19, 2016

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.

Model Creation

Load Data

Model 1

```
#dfraw <- read.csv(url("https://raw.githubusercontent.com/dsmilo/DATA621/master/HW1/data/moneyball-training-da
dfraw = trainingdata
dfremove <- subset(dfraw, TEAM_BATTING_SO == 0 | TEAM_PITCHING_SO == 0 | TEAM_BASERUN_SB == 0 | TEAM_BATTING_B
df <- subset(dfraw, !(INDEX %in% dfremove))
df1 <- df[, -c(1,10,11)] #Remove caught stealing and hit by pitcher variables
#View(df)
#View(df1)
#summary(df)

df$TEAM_BATTING_HSO <- df$TEAM_BATTING_H/df$TEAM_BATTING_SO #Ratio of hits to strikeouts

fit1 <- lm(TARGET_WINS~.-TEAM_PITCHING_HR-TEAM_BATTING_SO-TEAM_BATTING_H, df)#Non-significant predictors remov
summary(fit1)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ . - TEAM_PITCHING_HR - TEAM_BATTING_SO -
##     TEAM_BATTING_H, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.2751  -6.1830   0.1977   4.9095  23.2062
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      59.4925078 39.8824408   1.492 0.137569
## INDEX            -0.0002778  0.0008527  -0.326 0.744962
## TEAM_BATTING_2B   0.0259782  0.0302387   0.859 0.391451
## TEAM_BATTING_3B  -0.1059715  0.0778332  -1.362 0.175089
## TEAM_BATTING_HR   0.0890756  0.0260665   3.417 0.000786 ***
## TEAM_BATTING_BB  -0.3724819  0.5509648  -0.676 0.499894
## TEAM_BASERUN_SB   0.0360986  0.0284201   1.270 0.205697
## TEAM_BASERUN_CS  -0.0186777  0.0721563  -0.259 0.796053
## TEAM_BATTING_HBP  0.0794072  0.0496219   1.600 0.111337
## TEAM_PITCHING_H   0.0226968  0.0287785   0.789 0.431364
## TEAM_PITCHING_BB  0.4263230  0.5502906   0.775 0.439543
## TEAM_PITCHING_SO -0.0324158  0.0342181  -0.947 0.344769
## TEAM_FIELDING_E  -0.1750225  0.0414445  -4.223 3.86e-05 ***
## TEAM_FIELDING_DP -0.1007458  0.0367780  -2.739 0.006791 **
## TEAM_BATTING_HSO -0.4083421 25.5209070  -0.016 0.987252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.476 on 176 degrees of freedom
##   (2059 observations deleted due to missingness)
## Multiple R-squared:  0.5466, Adjusted R-squared:  0.5105
## F-statistic: 15.15 on 14 and 176 DF,  p-value: < 2.2e-16
```

```
step1 <- step(fit1)
```

```
## Start:  AIC=830.8
## 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 +
##     TEAM_BATTING_HSO) - TEAM_PITCHING_HR - TEAM_BATTING_SO -
##     TEAM_BATTING_H
##
##                    Df Sum of Sq   RSS    AIC
## - TEAM_BATTING_HSO  1      0.02 12644 828.80
## - TEAM_BASERUN_CS   1      4.81 12649 828.88
## - INDEX             1      7.63 12652 828.92
## - TEAM_BATTING_BB   1     32.84 12677 829.30
## - TEAM_PITCHING_BB  1     43.12 12688 829.45
## - TEAM_PITCHING_H   1     44.69 12689 829.48
## - TEAM_BATTING_2B   1     53.02 12697 829.60
## - TEAM_PITCHING_SO  1     64.47 12709 829.78
## - TEAM_BASERUN_SB   1    115.91 12760 830.55
## <none>                          12644 830.80
## - TEAM_BATTING_3B   1    133.18 12778 830.81
## - TEAM_BATTING_HBP  1    183.97 12828 831.56
## - TEAM_FIELDING_DP  1    539.09 13183 836.78
## - TEAM_BATTING_HR   1    838.95 13483 841.07
## - TEAM_FIELDING_E   1   1281.26 13926 847.24
##
## Step:  AIC=828.8
## TARGET_WINS ~ INDEX + 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_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP
##
##                    Df Sum of Sq   RSS    AIC
## - TEAM_BASERUN_CS   1      4.82 12649 826.88
## - INDEX             1      7.70 12652 826.92
## - TEAM_BATTING_BB   1     33.06 12677 827.30
## - TEAM_PITCHING_BB  1     43.40 12688 827.46
## - TEAM_BATTING_2B   1     53.22 12698 827.61
## - TEAM_BASERUN_SB   1    117.00 12761 828.56
## <none>                          12644 828.80
## - TEAM_BATTING_3B   1    134.40 12779 828.82
## - TEAM_BATTING_HBP  1    184.37 12829 829.57
## - TEAM_PITCHING_H   1    210.55 12855 829.96
## - TEAM_FIELDING_DP  1    539.50 13184 834.78
## - TEAM_BATTING_HR   1    855.30 13500 839.31
## - TEAM_FIELDING_E   1   1283.95 13928 845.28
## - TEAM_PITCHING_SO  1   1310.14 13954 845.64
##
## Step:  AIC=826.88
## TARGET_WINS ~ INDEX + TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR +
##     TEAM_BATTING_BB + TEAM_BASERUN_SB + TEAM_BATTING_HBP + TEAM_PITCHING_H +
##     TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                    Df Sum of Sq   RSS    AIC
## - INDEX             1      6.74 12656 824.98
## - TEAM_BATTING_BB   1     33.43 12683 825.38
## - TEAM_PITCHING_BB  1     43.86 12693 825.54
## - TEAM_BATTING_2B   1     52.37 12702 825.67
## <none>                          12649 826.88
## - TEAM_BASERUN_SB   1    140.51 12790 826.99
## - TEAM_BATTING_3B   1    145.53 12795 827.06
## - TEAM_BATTING_HBP  1    183.89 12833 827.63
## - TEAM_PITCHING_H   1    219.61 12869 828.16
## - TEAM_FIELDING_DP  1    547.06 13196 832.96
## - TEAM_BATTING_HR   1    868.68 13518 837.56
## - TEAM_PITCHING_SO  1   1305.39 13955 843.64
## - TEAM_FIELDING_E   1   1383.12 14032 844.70
##
## Step:  AIC=824.98
## TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR +
##     TEAM_BATTING_BB + TEAM_BASERUN_SB + TEAM_BATTING_HBP + TEAM_PITCHING_H +
##     TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                    Df Sum of Sq   RSS    AIC
## - TEAM_BATTING_BB   1     29.39 12685 823.42
## - TEAM_PITCHING_BB  1     39.34 12695 823.57
## - TEAM_BATTING_2B   1     51.40 12707 823.75
## <none>                          12656 824.98
## - TEAM_BASERUN_SB   1    140.94 12797 825.09
## - TEAM_BATTING_3B   1    143.47 12799 825.13
## - TEAM_BATTING_HBP  1    179.55 12836 825.67
## - TEAM_PITCHING_H   1    222.44 12878 826.31
## - TEAM_FIELDING_DP  1    581.74 13238 831.56
## - TEAM_BATTING_HR   1    894.01 13550 836.02
## - TEAM_PITCHING_SO  1   1312.45 13968 841.83
## - TEAM_FIELDING_E   1   1376.72 14033 842.70
##
## Step:  AIC=823.42
## TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR +
##     TEAM_BASERUN_SB + TEAM_BATTING_HBP + TEAM_PITCHING_H + TEAM_PITCHING_BB +
##     TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                    Df Sum of Sq   RSS    AIC
## - TEAM_BATTING_2B   1     46.94 12732 822.13
## <none>                          12685 823.42
## - TEAM_BASERUN_SB   1    142.30 12828 823.55
## - TEAM_BATTING_3B   1    149.66 12835 823.66
## - TEAM_BATTING_HBP  1    181.78 12867 824.14
## - TEAM_PITCHING_H   1    224.17 12910 824.77
## - TEAM_FIELDING_DP  1    601.89 13287 830.28
## - TEAM_BATTING_HR   1    875.98 13561 834.18
## - TEAM_PITCHING_SO  1   1305.57 13991 840.13
## - TEAM_FIELDING_E   1   1353.41 14039 840.78
## - TEAM_PITCHING_BB  1   2317.51 15003 853.47
##
## Step:  AIC=822.13
## TARGET_WINS ~ TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BASERUN_SB +
##     TEAM_BATTING_HBP + TEAM_PITCHING_H + TEAM_PITCHING_BB + TEAM_PITCHING_SO +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP
##
##                    Df Sum of Sq   RSS    AIC
## - TEAM_BASERUN_SB   1    108.69 12841 821.75
## <none>                          12732 822.13
## - TEAM_BATTING_3B   1    158.21 12890 822.49
## - TEAM_BATTING_HBP  1    185.64 12918 822.89
## - TEAM_PITCHING_H   1    494.78 13227 827.41
## - TEAM_FIELDING_DP  1    620.16 13352 829.21
## - TEAM_BATTING_HR   1    839.55 13572 832.32
## - TEAM_PITCHING_SO  1   1259.19 13992 838.14
## - TEAM_FIELDING_E   1   1399.47 14132 840.05
## - TEAM_PITCHING_BB  1   2358.84 15091 852.59
##
## Step:  AIC=821.75
## TARGET_WINS ~ TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_HBP +
##     TEAM_PITCHING_H + TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP
##
##                    Df Sum of Sq   RSS    AIC
## <none>                          12841 821.75
## - TEAM_BATTING_3B   1    135.35 12976 821.75
## - TEAM_BATTING_HBP  1    176.14 13017 822.35
## - TEAM_PITCHING_H   1    577.33 13418 828.15
## - TEAM_FIELDING_DP  1    732.75 13574 830.35
## - TEAM_BATTING_HR   1    752.59 13594 830.63
## - TEAM_PITCHING_SO  1   1249.14 14090 837.48
## - TEAM_FIELDING_E   1   1335.48 14176 838.65
## - TEAM_PITCHING_BB  1   2364.62 15206 852.03
```

```
summary(step1)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_3B + TEAM_BATTING_HR +
##     TEAM_BATTING_HBP + TEAM_PITCHING_H + TEAM_PITCHING_BB + TEAM_PITCHING_SO +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.562  -5.939   0.031   5.255  21.696
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      58.241228  19.168933   3.038  0.00273 **
## TEAM_BATTING_3B  -0.104216   0.075242  -1.385  0.16773
## TEAM_BATTING_HR   0.080436   0.024628   3.266  0.00130 **
## TEAM_BATTING_HBP  0.077262   0.048899   1.580  0.11584
## TEAM_PITCHING_H   0.030486   0.010657   2.861  0.00472 **
## TEAM_PITCHING_BB  0.054826   0.009470   5.789 3.04e-08 ***
## TEAM_PITCHING_SO -0.030616   0.007276  -4.208 4.05e-05 ***
## TEAM_FIELDING_E  -0.172105   0.039558  -4.351 2.26e-05 ***
## TEAM_FIELDING_DP -0.113640   0.035263  -3.223  0.00151 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.4 on 182 degrees of freedom
##   (2059 observations deleted due to missingness)
## Multiple R-squared:  0.5395, Adjusted R-squared:  0.5193
## F-statistic: 26.66 on 8 and 182 DF,  p-value: < 2.2e-16
```

```
#Correlation Matrix
#View(round(cor(df1),2))

#These are variables that I tried but didn't turn out to be valuable
df1$TEAM_BATTING_1B <- df1$TEAM_BATTING_H - df1$TEAM_BATTING_2B - df1$TEAM_BATTING_3B - df1$TEAM_BATTING_HR #S
df1$TEAM_BATTING_HRP <- df1$TEAM_BATTING_HR/df1$TEAM_BATTING_H #Home runs as a percentage of base hits
```

Imputing Missing values

```
## Loading required package: Rcpp
```

```
## mice 2.25 2015-11-09
```

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

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

```
##
## Call:
## lm(formula = TARGET_WINS ~ . - INDEX, data = trainingdata)
##
## 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 = trainingdata)
##
## 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
```

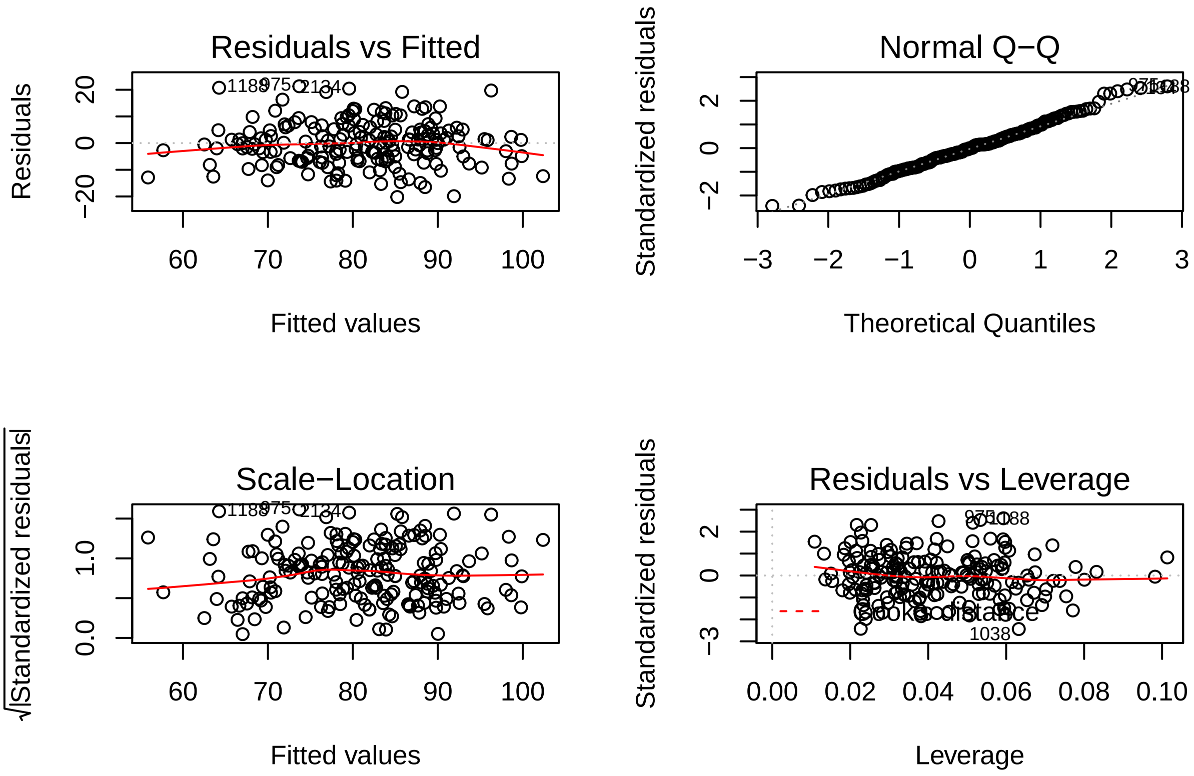
```
#####Generate predictions using the stepFull model
predictionsStepFull <- predict(stepFull, trainingDataRaw)
#View(predictionsStepFull)
```

Generate the RMSE of the stepFull model

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

```
trainingDataRaw = trainingdata
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
```

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

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

Model Selection and Prediction