Data 621 Homework 1

jim lung

02-16-2018

Introduction

We will explore, analyze and model a data set representing a professional baseball team from the years 1871 to 2006 inclusive. Each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season. We will build three multiple linear regression models on the training data to predict the number of wins for the team.

Data Exploration

The training data set consists of 2276 records. Each record represents the performance of a team during a one year baseball season. The response variable, which is what we want to train our models to predict, is "TARGET_WINS."

The predictor variables listed below represent the number of batting, running and fielding events that occurred during games. The events were captured because they are posited to have either a positive (denoted by +) or negative (-) impact on winning the game.

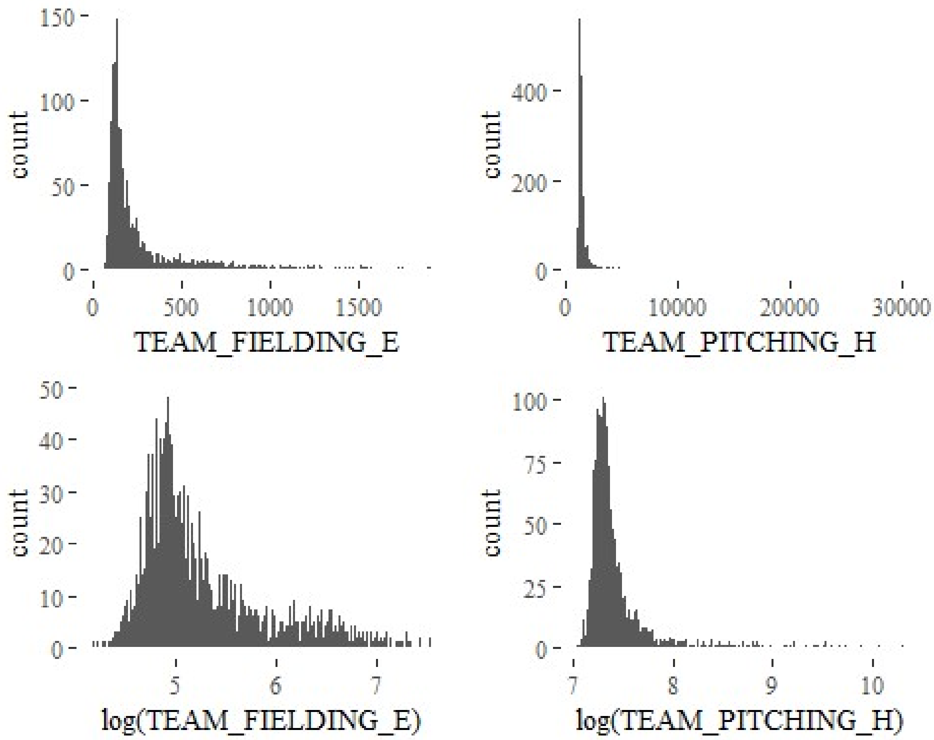
Variable Name	Impact	Definition
TEAM_BATTING_H	+	Base Hits by batters (1B,2B,3B,HR)
TEAM_BATTING_2B	+	Doubles by batters (2B)
TEAM_BATTING_3B	+	Triples by batters (3B)
TEAM_BATTING_HR	+	Homeruns by batters (4B)
TEAM_BATTING_BB	+	Walks by batters
TEAM_BATTING_HBP	+	Batters hit by pitch
TEAM_BATTING_SO	-	Strikeouts by batters
TEAM_BASERUN_SB	+	Stolen bases
TEAM_BASERUN_CS	-	Caught stealing
TEAM_FIELDING_E	-	Errors
TEAM_FIELDING_DP	+	Double Plays
TEAM_PITCHING_BB	-	Walks allowed
TEAM_PITCHING_H	-	Hits allowed

TEAM_PITCHING_HR	-	Homeruns allowed
TEAM_PITCHING_SO	+	Strikeouts by pitchers

Data Preparation

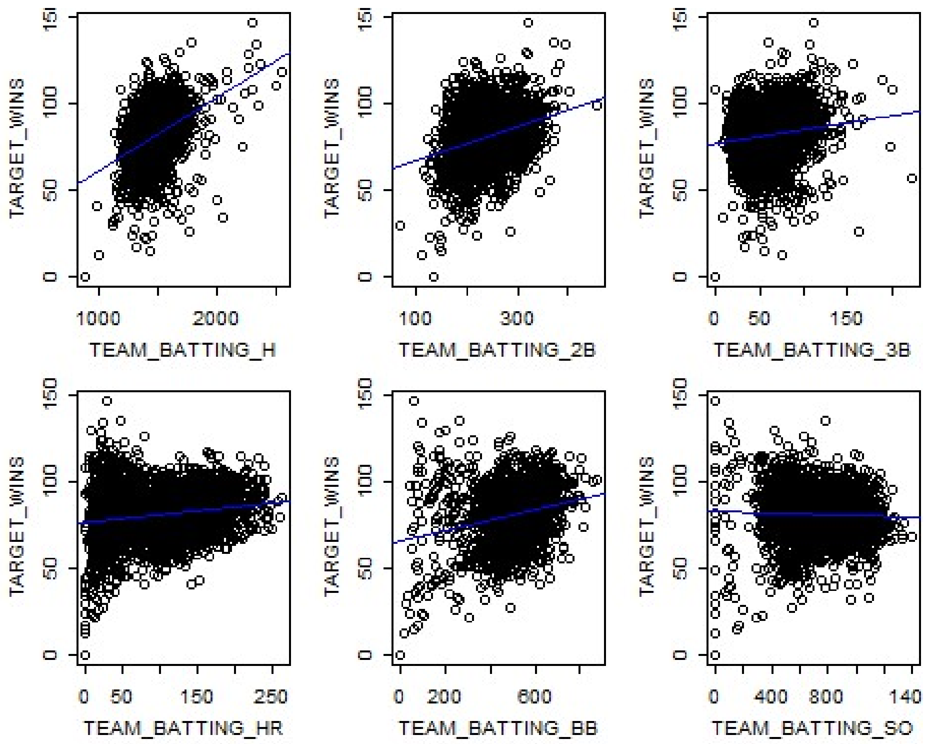
Most of the rows of data are missing at least one value, and one variable, TEAM_BATTING_HBP, is so sparsely populated (191 of 2276 records) that we will exclude it from consideration altogether. For the other variables, we first try filling in the average value of the data field in the missing records. This diminishes the signal in the data, as the overall r-squared value (the variation explained by the model) drops from 0.44 to 0.32. We could use a regression model to fill in the blanks but since we would then be using that filled in data for further regression analysis, in my view, it would not add predictive value. Since we do have a very large number of complete records, excluding TEAM_BATTING_HBP, it may better to simply ignore incomplete records when evaluating the significance and value of each variable, rather than filling in with 'dummy' data.

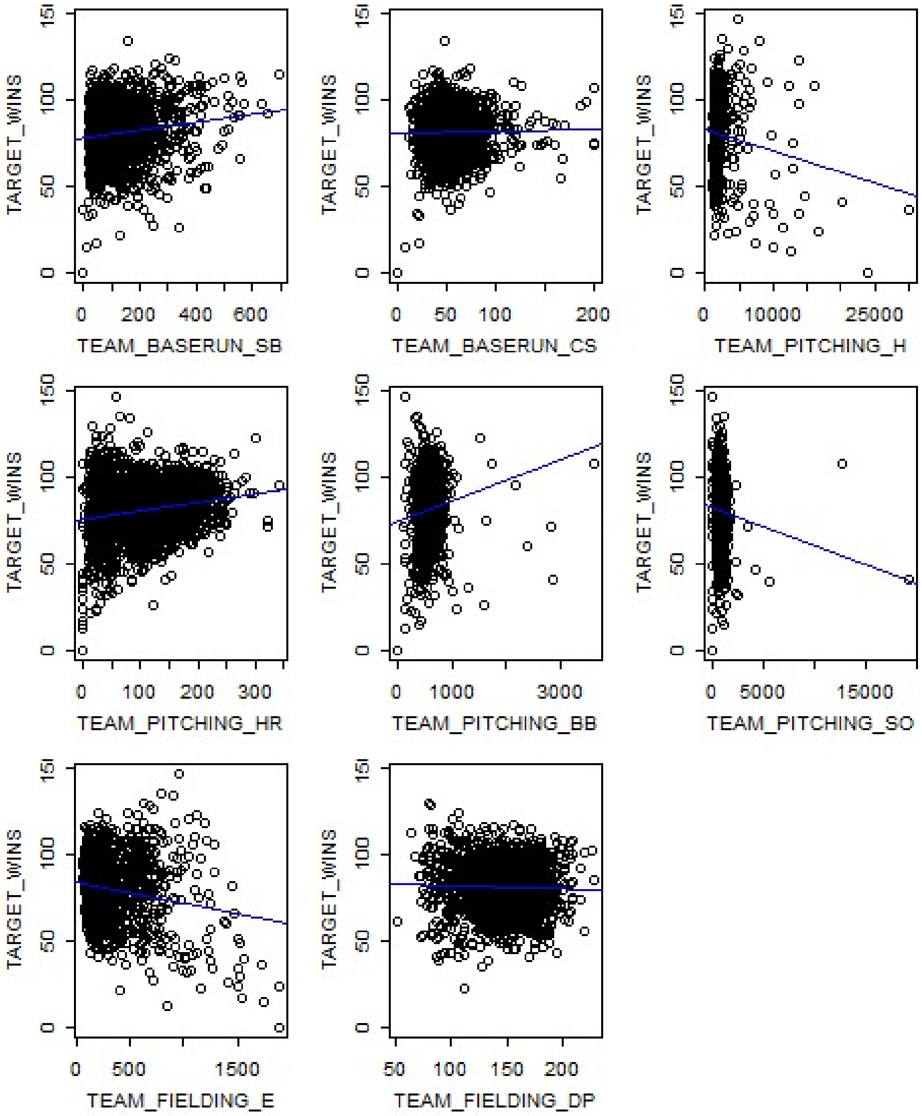
Number of errors and hits allowed, for example, had some outliers, but a histogram of the data shows that it follows a non-normal, very right skewed distribution:



These two variables look more normal when a log transformation is applied to them. Hits allowed, after a log transformation is performed, seems to still have outliers.

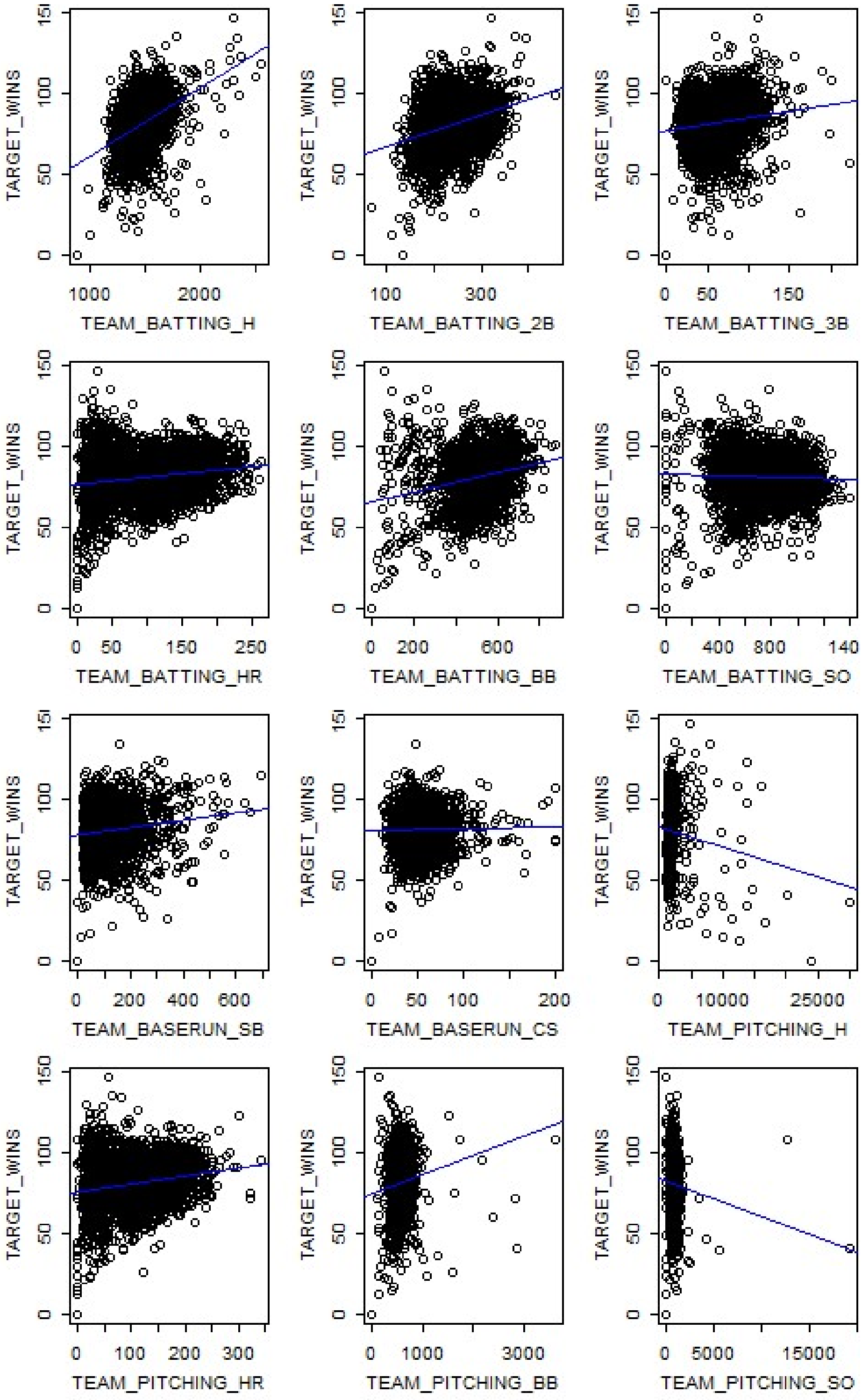
Below are scatterplots of each of the predictive attributes. We will look at these for outliers or other structure indicating non-normality.

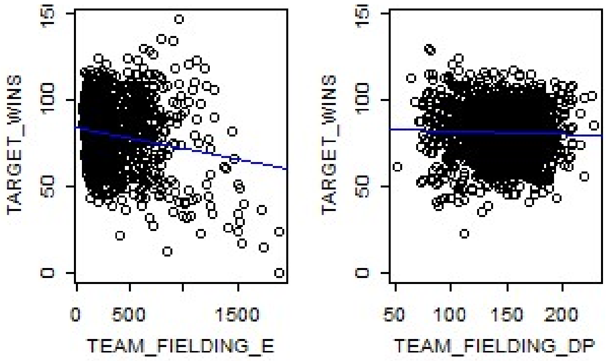




There are a small number of extreme outliers in TEAM_PITCHING_H, TEAM_PITCHING_BB and TEAM_PITCHING_SO that have an outsize effect on the model, we will remove those as they are most likely observational errors.

Also, since TEAM_BATTING_H is a combination of TEAM_BATTING_2B, TEAM_BATTING_3B, TEAM_BATTING_HR (and also includes batted singles), we will create a new variable TEAM_BATTING_1B equaling TEAM_BATTING_H - TEAM_BATTING_2B - TEAM_BATTING_3B - TEAM_BATTING_HR, just to see if there is any significance in hitting singles versus any successful hit.


TARGET_WINS
0
50
100
150
1000
2000
TEAM_BATTING_H
100
300
TEAM_BATTING_2B
0
50
150
TEAM_BATTING_3B
0
50
150
250
TEAM_BATTING_HR
0
200
600
TEAM_BATTING_BB
0
400
800
140
TEAM_BATTING_SO
0
200
400
600
TEAM_BASERUN_SB
0
50
100
200
TEAM_BASERUN_CS
0
10000
25000
TEAM_PITCHING_H
0
100
200
300
TEAM_PITCHING_HR
0
1000
3000
TEAM_PITCHING_BB
0
5000
15000
TEAM_PITCHING_SO

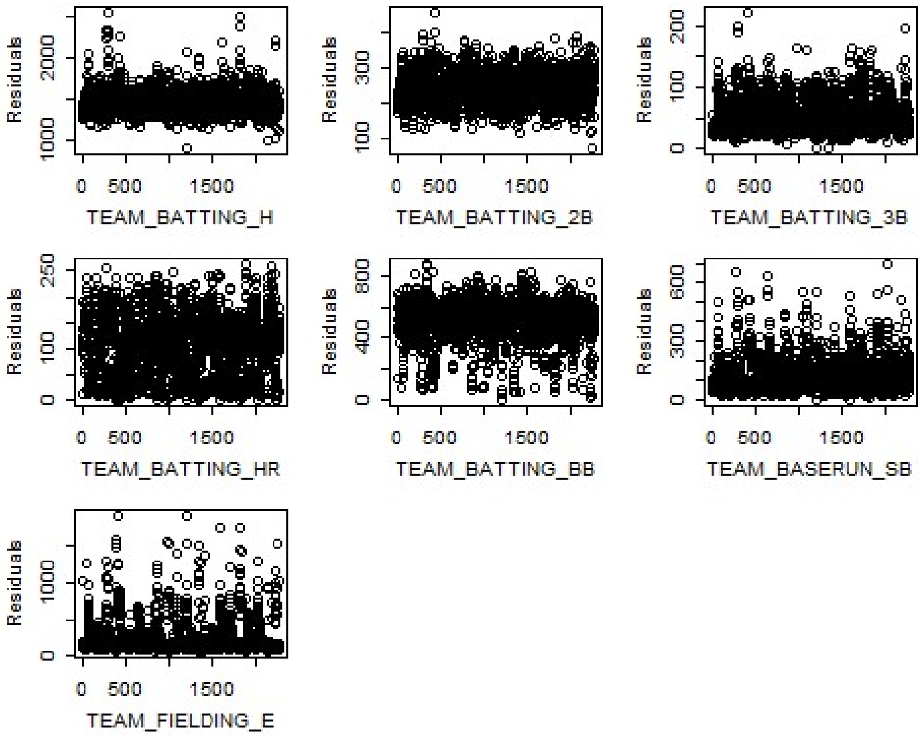


The A-B line for batting singles is similar to the other batting A-B lines, nothing new to see here. Looking at the cleansed data, the pitching data now seems complementary to the hitting data, (as it should since they are representing the same events from two perspectives), so we will ignore the pitching data.

Next we will look at the key attributes of the individual linear models for each variable. We will make a preliminary decision whether or not to include the remaining variables based on the Coefficient (impact on wins), Error (trustworthiness of Coefficient), R-Squared (explanatory value) and P-Value (significance).

Attribute	Coefficient	Error	Error Percent	R_Squared	P-Value	Use
TEAM_BATTING_H	0.042	0.002	5	0.151	0.0000	Yes
TEAM_BATTING_2B	0.097	0.007	7	0.083	0.0000	Yes
TEAM_BATTING_3B	0.080	0.012	15	0.020	0.0000	Yes
TEAM_BATTING_HR	0.046	0.005	11	0.031	0.0000	Yes
TEAM_BATTING_BB	0.030	0.003	10	0.054	0.0000	Yes
TEAM_BATTING_SO	-0.002	0.001	50	0.001	0.1389	No
TEAM_BASERUN_SB	0.023	0.004	17	0.018	0.0000	Yes
TEAM_BASERUN_CS	0.013	0.015	115	0.000	0.3853	No
TEAM_BATTING_HBP	0.069	0.068	99	0.000	0.3122	No
TEAM_PITCHING_H	-0.001	0.000	0	0.012	0.0000	No
TEAM_PITCHING_HR	0.049	0.005	10	0.035	0.0000	No
TEAM_PITCHING_BB	0.012	0.002	17	0.015	0.0000	No
TEAM_PITCHING_SO	-0.002	0.001	50	0.006	0.0003	No
TEAM_FIELDING_E	-0.012	0.001	8	0.031	0.0000	Yes
TEAM_FIELDING_DP	-0.019	0.012	63	0.001	0.1201	No

Below are residual plots for the remaining variables. A lack of structure in a residual plot indicates near constant variance.



Finally we should check for collinearity among variables, by measuring the Variance Inflation Factor for each variable.

```
##  TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_HR
##        2.982781        2.455073        3.082753        2.727518
## TEAM_BATTING_BB TEAM_BASERUN_SB TEAM_FIELDING_E
##        1.408649        1.631054        2.265417
```

This analysis suggests that the TEAM_BATTING_H variable is the highly redundant, as is TEAM_FIELDING_E. Since TEAM_BATTING_H is composed of four other variables, let's remove it first and check for redundancy again.

```
## TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_HR TEAM_BATTING_BB
##        1.364003        2.314762        2.589708        1.408180
## TEAM_BASERUN_SB TEAM_FIELDING_E
##        1.607874        1.978712
```

Now all the remaining variables have a low VIF value, and we have satisfied all the requirements for removing unsuitable variables from our multiple linear regression model.

Build Models

Model 1

For our first model, we will use TEAM_BATTING_2B, TEAM_BATTING_3B and TEAM_BATTING_HR.

Batting is good for winning, particularly batting doubles and triples and home runs. This is seen by the strong positive coefficents (Estimate), low standard errors / p-value, and reasonably high R-squared value in the linear model for these variables.

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR, data = dfa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.045  -9.180   0.669   9.549  54.576
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     45.528985   1.807448   25.19   <2e-16 ***
## TEAM_BATTING_2B  0.061838   0.007388    8.37   <2e-16 ***
## TEAM_BATTING_3B  0.211353   0.014431   14.64   <2e-16 ***
## TEAM_BATTING_HR  0.087001   0.007354   11.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.4 on 2272 degrees of freedom
## Multiple R-squared:  0.1655, Adjusted R-squared:  0.1644
## F-statistic: 150.2 on 3 and 2272 DF,  p-value: < 2.2e-16
```

Below is a listing of the first six records of the evaluation data:

```
##   INDEX TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_
HR
## 1     9           1209             170              33
83
## 2    10           1221             151              29
88
## 3    14           1395             183              29
93
## 4    47           1539             309              29             1
59
## 5    60           1445             203              68
 5
## 6    63           1431             236              53
10
```

```
##   TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB TEAM_BASERUN_CS
## 1             447            1080              62              50
## 2             516             929              54              39
## 3             509             816              59              47
## 4             486             914             148              57
## 5              95             416              NA              NA
## 6             215             377              NA              NA
##   TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
## 1               NA            1209               83              447
## 2               NA            1221               88              516
## 3               NA            1395               93              509
## 4               42            1539              159              486
## 5               NA            3902               14              257
## 6               NA            2793               20              420
##   TEAM_PITCHING_SO TEAM_FIELDING_E TEAM_FIELDING_DP
## 1             1080             140              156
## 2              929             135              164
## 3              816             156              153
## 4              914             124              154
## 5             1123             616              130
## 6              736             572              105
```

Using the Predict function, we can predict the number of wins for each evaluation data record.

```
myModel1_Predictions <- predict.lm(myModel1,dfEval) #predict
head(myModel1_Predictions)
```

```
##        1        2        3        4        5        6
## 70.23722 68.65189 71.06571 84.59940 72.88912 72.19449
```

- **Model 2**

In this model we use all the Batting variables. The adjusted R-squared value increases markedly, which should yield much better predictions than the first model. This model illustrates the value of batting singles and earning walks.

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB, data = dfa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.408  -8.600   0.516   9.137  55.280
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.321537   3.466082   0.958    0.338
```

```
## TEAM_BATTING_H   0.037463   0.003075  12.182  < 2e-16 ***
## TEAM_BATTING_2B -0.007777   0.009018  -0.862    0.389
## TEAM_BATTING_3B  0.098673   0.016400   6.017 2.07e-09 ***
## TEAM_BATTING_HR  0.048973   0.007752   6.318 3.19e-10 ***
## TEAM_BATTING_BB  0.027859   0.002805   9.932  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.79 on 2270 degrees of freedom
## Multiple R-squared:  0.2356, Adjusted R-squared:  0.2339
## F-statistic: 139.9 on 5 and 2270 DF,  p-value: < 2.2e-16
```

Here are sample predictions for the second model:

```
myModel2_Predictions <- predict.lm(myModel2,dfEval) #predict
head(myModel2_Predictions)

##        1        2        3        4        5        6
## 67.06559 69.43531 75.75481 82.76104 65.47753 66.80424
```

Model 3

In this model we use all the variables we determined above to be useful. The adjusted R-squared value again increases substantially while the residual standard error drops incrementally. This model illustrates the incremental value of stolen bases, and the negative impact of fielding errors.

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BASER
UN_SB +
##     TEAM_FIELDING_E, data = dfa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.935  -8.318   0.001   8.066  49.247
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.306426   3.419753   0.382    0.702
## TEAM_BATTING_H   0.050507   0.003290  15.352  < 2e-16 ***
## TEAM_BATTING_2B -0.050423   0.008955  -5.631 2.03e-08 ***
## TEAM_BATTING_3B  0.078646   0.017009   4.624 3.99e-06 ***
## TEAM_BATTING_HR  0.043230   0.007342   5.888 4.52e-09 ***
## TEAM_BATTING_BB  0.021012   0.003157   6.655 3.60e-11 ***
## TEAM_BASERUN_SB  0.046990   0.003809  12.337  < 2e-16 ***
## TEAM_FIELDING_E -0.037125   0.002322 -15.988  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.12 on 2137 degrees of freedom
##   (131 observations deleted due to missingness)
## Multiple R-squared:  0.328,  Adjusted R-squared:  0.3258
## F-statistic:   149 on 7 and 2137 DF,  p-value: < 2.2e-16
```

Here are sample predictions for the third model:

```
myModel3_Predictions <- predict(myModel3,dfEval, interval='confidence')
 #predict
head(myModel3_Predictions)
```

```
##        fit      lwr      upr
## 1 67.08831 65.85675 68.31987
## 2 69.81350 68.45354 71.17345
## 3 76.51249 75.22475 77.80022
## 4 85.17213 83.89760 86.44666
## 5       NA       NA       NA
## 6       NA       NA       NA
```

Select Models

Model 3 is the best multiple linear regression model because it uses all of the relevant available information to provide the strongest estimate. It has the highest Adjusted R-squared value (0.33) and the lowest p-value (~0). However, in cases where not all variables are present in Model 3, we should use Model 2.

Predictions from Model 3 are shown directly above. The residuals plots for Model 3 are shown near the end of the Data Preparation section, where issues of collinearity were resolved (colinear variables were eliminated).

A topic for further study would be to develop a virtual model that seamlessly switches between the two models as needed, without corrupting either model with imputed data.

Conclusions

This model would benefit from more conceptual analysis of what these measures mean, and more analysis of how this compares to the observed relationships between the variables.

It's pretty clear that there should be relationships betweeen base hits, singles, doubles, triples, and home runs, but looking at pairwise correlation, only some pairs showed correlation. There’s no clear reason why base hits should be correlated to doubles, but not triples or home runs for example.