Moneyball - CUNY Data Science 621

Eric Hirsch

2/20/2021

Description of the Dataset

a. ASSIGNMENT: In this assignment we explore, analyze and model a data set containing approximately 2276 records, each 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 a multiple linear regression model on the training data to predict the number of wins for the team.

b. THE ISSUE OF HIDDEN GROUPINGS: An issue with the data is hidden groupings. Records may not be independent of each other, as team data in one year will be related to team data in the next year. We know that if some records were adjusted to match a longer season, there may be an "eras of baseball" effect as teams from earlier years behave differently from later ones. Finally, within the record, columns may not be independent. In particular, teams with high offensive stats (like hitting) may have lower defensive stats (like pitching), as the teams on limited budgets make strategic choices between the two. We will attempt to address some of these issues in this analysis.

1. Data Exploration

All of the columns in the dataset are numeric. We begin by examining their means, medians and distributions.

```
##      INDEX         TARGET_WINS      BATTING_H      BATTING_2B
##  Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0
##  1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0
##  Median :1270.5   Median : 82.00   Median :1454   Median :238.0
##  Mean   :1268.5   Mean   : 80.79   Mean   :1469   Mean   :241.2
##  3rd Qu.:1915.5   3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0
##  Max.   :2535.0   Max.   :146.00   Max.   :2554   Max.   :458.0
##
##    BATTING_3B       BATTING_HR       BATTING_BB      BATTING_SO
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.0   Min.   :   0.0
##  1st Qu.: 34.00   1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0
##  Median : 47.00   Median :102.00   Median :512.0   Median : 750.0
##  Mean   : 55.25   Mean   : 99.61   Mean   :501.6   Mean   : 735.6
##  3rd Qu.: 72.00   3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0
##  Max.   :223.00   Max.   :264.00   Max.   :878.0   Max.   :1399.0
##                                                    NA's   :102
##    BASERUN_SB      BASERUN_CS     BATTING_HBP      PITCHING_H
##  Min.   :  0.0   Min.   :  0.0   Min.   :29.00   Min.   : 1137
```

```
##  1st Qu.: 66.0   1st Qu.: 38.0   1st Qu.:50.50   1st Qu.: 1419
##  Median :101.0   Median : 49.0   Median :58.00   Median : 1518
##  Mean   :124.8   Mean   : 52.8   Mean   :59.36   Mean   : 1779
##  3rd Qu.:156.0   3rd Qu.: 62.0   3rd Qu.:67.00   3rd Qu.: 1682
##  Max.   :697.0   Max.   :201.0   Max.   :95.00   Max.   :30132
##  NA's   :131     NA's   :772     NA's   :2085
##   PITCHING_HR     PITCHING_BB      PITCHING_SO        FIELDING_E
##  Min.   :  0.0   Min.   :   0.0   Min.   :    0.0   Min.   :  65.0
##  1st Qu.: 50.0   1st Qu.: 476.0   1st Qu.:  615.0   1st Qu.: 127.0
##  Median :107.0   Median : 536.5   Median :  813.5   Median : 159.0
##  Mean   :105.7   Mean   : 553.0   Mean   :  817.7   Mean   : 246.5
##  3rd Qu.:150.0   3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2
##  Max.   :343.0   Max.   :3645.0   Max.   :19278.0   Max.   :1898.0
##                                   NA's   :102
##   FIELDING_DP
##  Min.   : 52.0
##  1st Qu.:131.0
##  Median :149.0
##  Mean   :146.4
##  3rd Qu.:164.0
##  Max.   :228.0
##  NA's   :286
```

We note that a number of columns have NAs. Batting_SO and Pitching_SO have the same number of NA's and may be related.

We more closely examine the distribution of columns in the dataset (fig. 1):

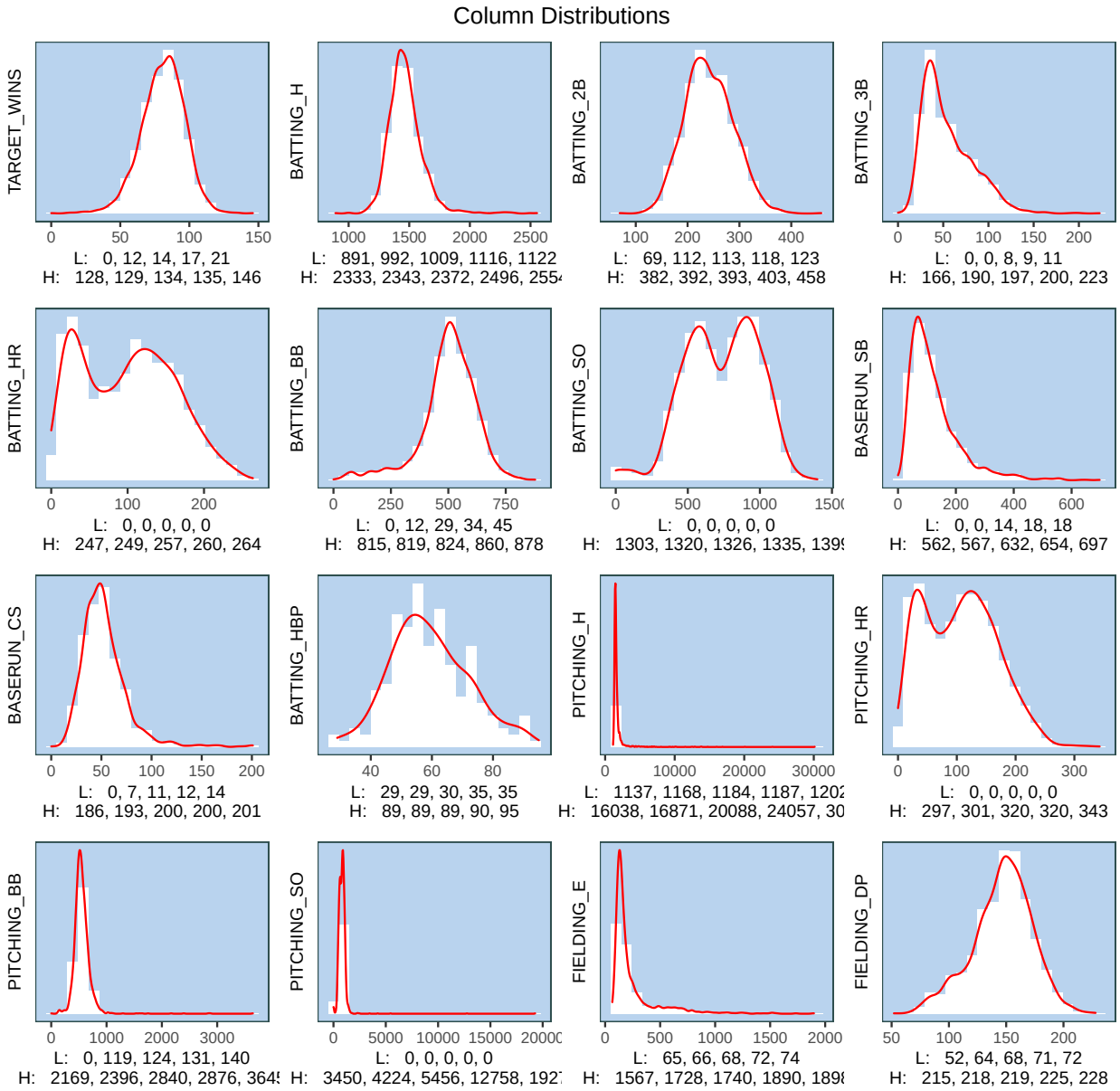


Fig. 1

Our dependent variable (Target Wins) appears to be normally distributed. However, a number of columns are severely skewed (Errors, Strikeouts, Pitching_H, etc.) A few columns (Batting SO, Pitching_HR and Batting_HR) have a bimodal distribution. This might point to some hidden groupings in the dataset.

Boxplots help us identify outliers (fig. 2):

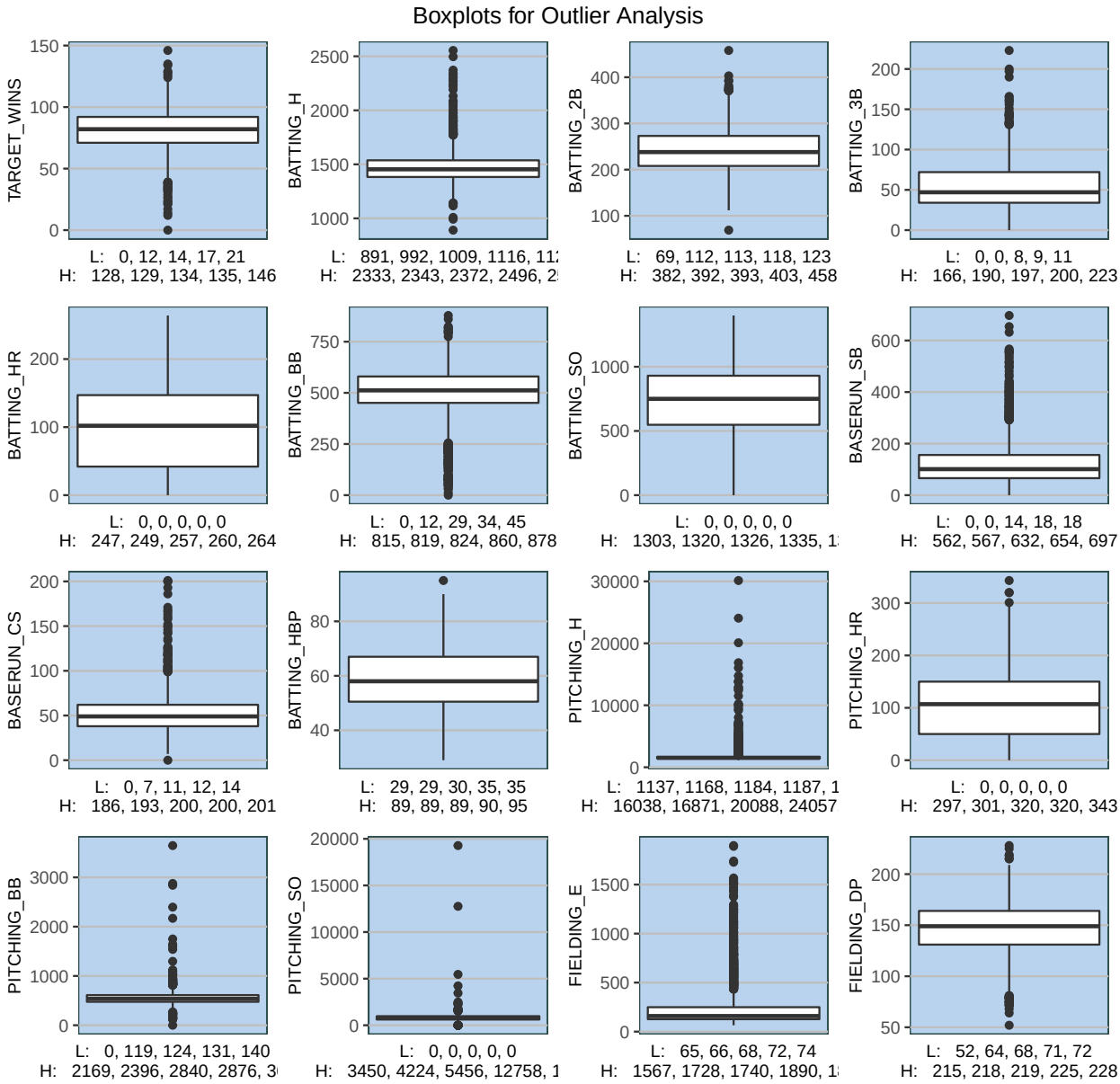


Fig. 2

There a number of outliers, both high and low. For example, there are many zeros, which may be implausible. In addition, many of the ranges appear extreme, such as giving up between 3,500 hits and 19,000 hits, or getting from 12 to over 800 walks.

We investigate correlations in the dataset, both between the dependent variable and the other variables (fig. 3), and between the dependent variables and each other (fig. 4).

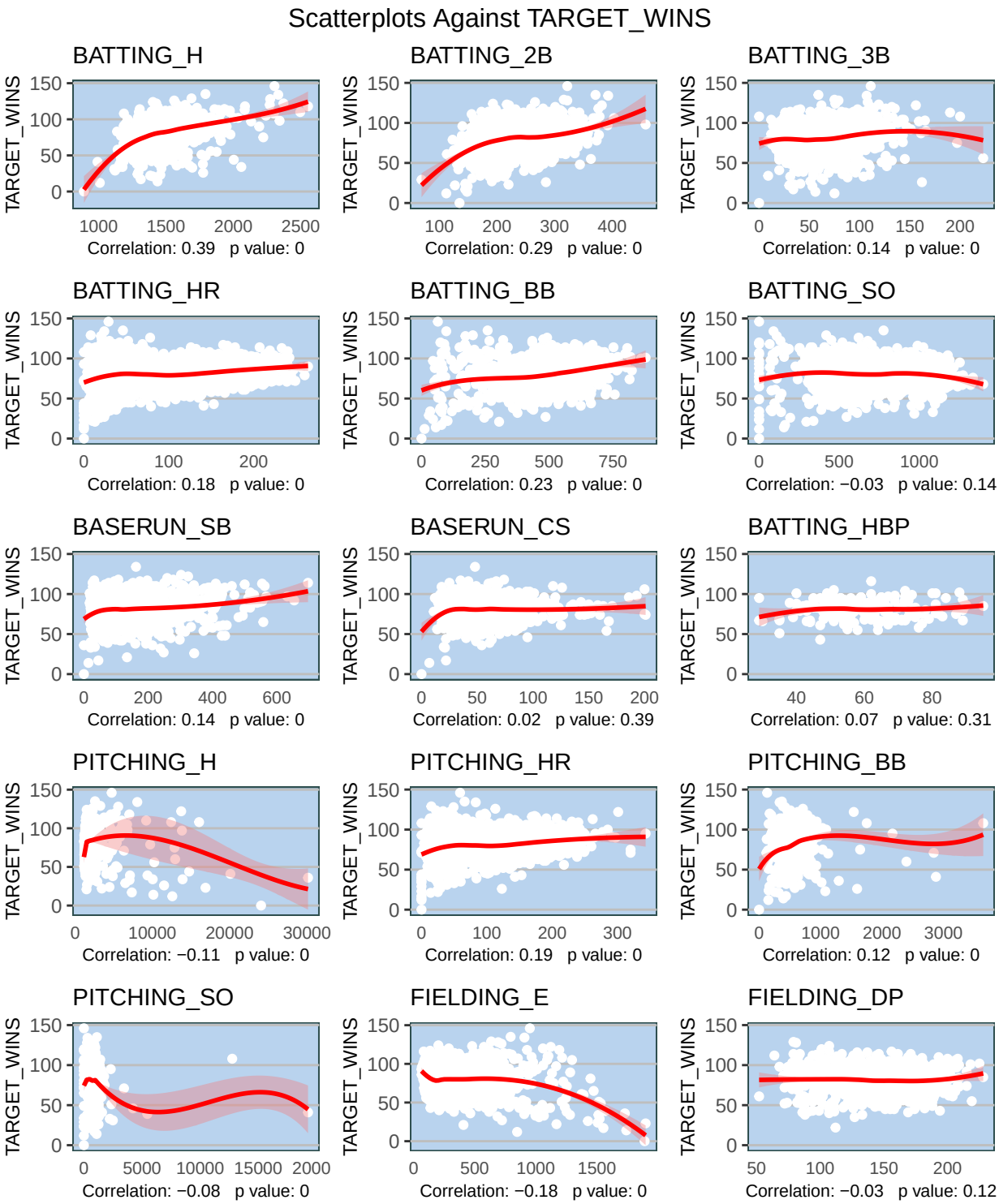


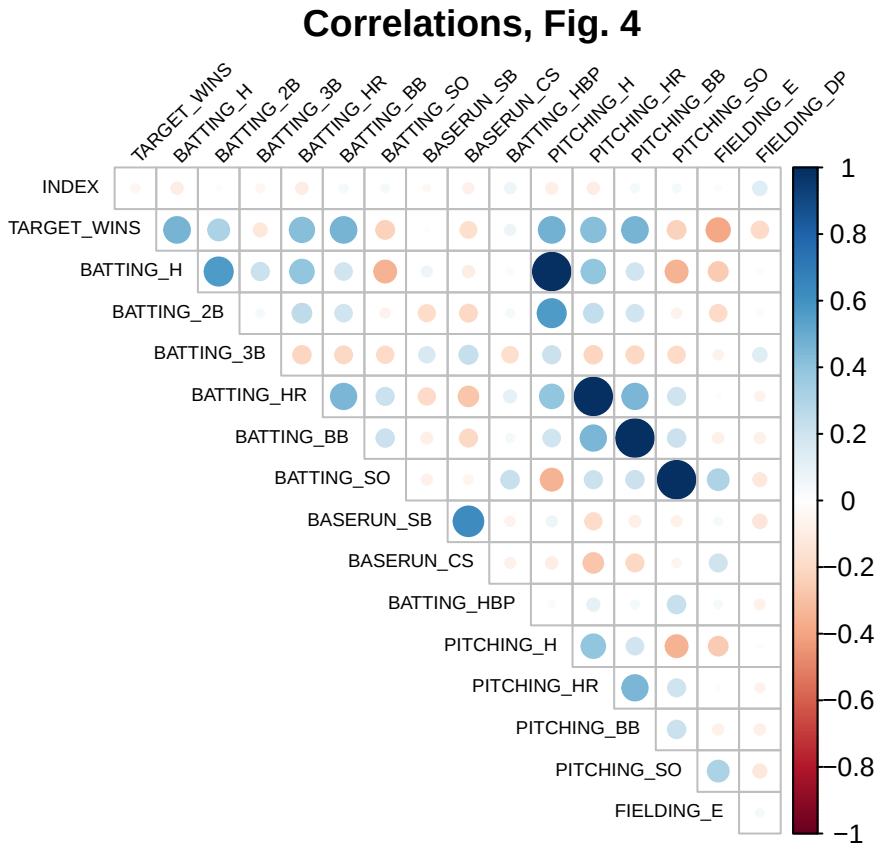
Fig.3

Here we see a number of puzzles, mainly among the pitching correlations. Hits should show a much stronger negative correlation, and in fact appear positive for a portion. Making double plays is surprisingly neutral, as are strikeouts. Pitching_HR is also positive when we would expect negative.

We do need to acknowledge here the possibility of strategy groupings (defense and offense) which may contribute to these anomalies. In other words, a team with poor pitching may have strong hitting, which

then wins games.

We can look for evidence of this possibility by examining multicollinearity:



Indeed, the pitching categories are strongly correlated with their hitting counterparts. All four of the pitching categories follow this pattern.

2. Data Preparation

We begin by devising a strategy for the NAs. We can eliminate the Batting_HBP and Baserun_CS columns because they have too many NA's. We also create flags for the other columns with significant NA's.

We are particularly interested in the SO columns because they do not appear random, and investigation establishes that they have complete overlap with each other. Fielding_DP and BR_SB aslo have some overlap. These may relate to eras of baseball when certain statistics were not collected. (see Fig. 5)

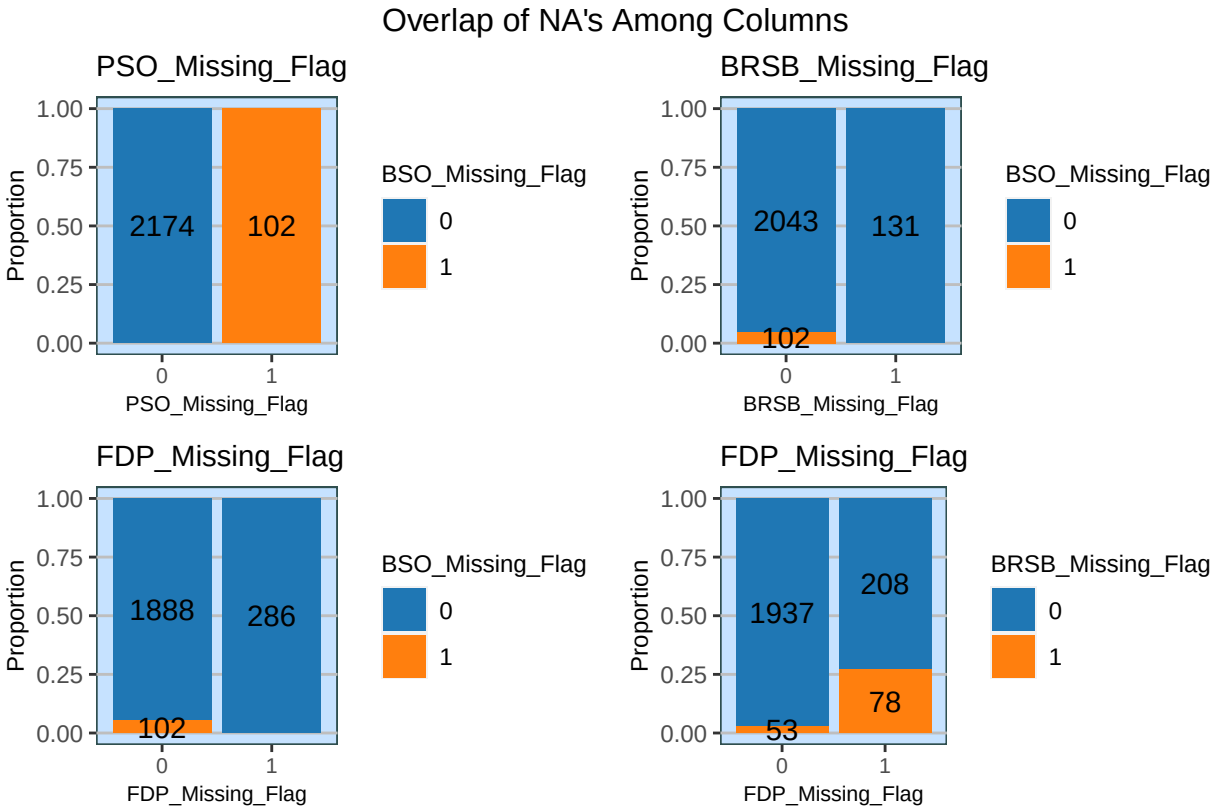


Fig. 5

We eliminate the pitching SO column because it is redundant. While not MCAR (missing completely at random), if the Batting_SO column is MAR (missing at random), we may be bale to eliminate these rows, as there are not so many (5% of the total).

One way to investigate the randomness of this missing cohort is to look for interactions between the cohort and other dataset columns. In fact, we see that there are a number of columns with strong, even extreme interactions (see fig. 6).

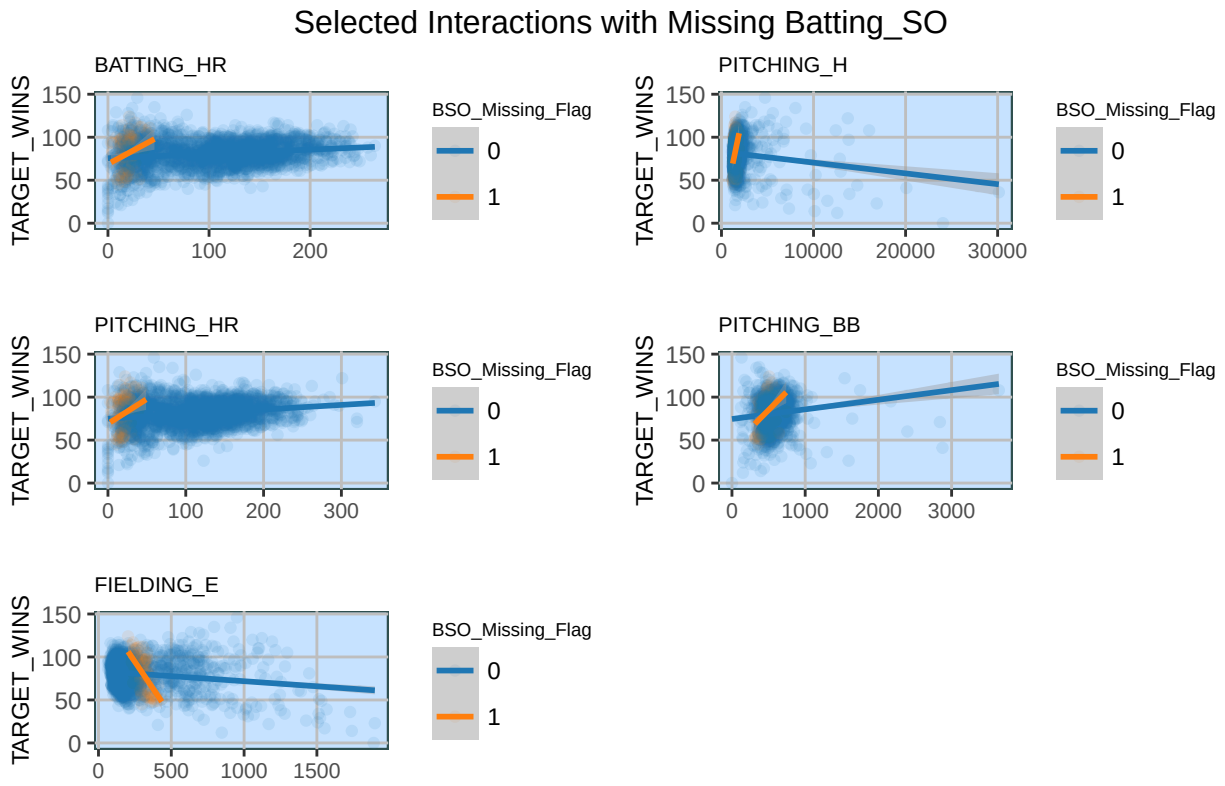


Fig. 6

It is possible this cohort represents a different baseball era when such statistics were not collected. In any case, we cannot eliminate these rows without losing critical data, so we employ the following strategy: 1) retain the rows and impute a value, 2) create a "missing" flag to keep track of the cohort, and 2) add interaction terms where appropriate.

Before we address imputation, we want to work with the implausible zeros in the dataet. In particular, we note that the 0s in Pitching_SO and Batting_SO are a complete overlap, and we can see from the histograms that the jump between 0 and the next lowest value is not smooth, and so we will treat them as NA's. We do the same with batting and pitching HR, since there is also a jump up after zero which suggests it is being used as an indicator of missing value.

Just so we have some reasonable criteria for imputation strategy, we compare the r-squared of three regressions - with NA's imputed as means, with NA's imputed as medians, and with NA rows eliminated altogether.

```
## [1] "type:" "mean"
## [1] "r2mean:" "0.4031"
## [1] "r2median:" "0.403"
## [1] "r2omit" "0.4019"
```

The mean and median have the same r-squared, while the elimination of the rows has a smaller r-squared. We therefore choose to impute the mean.

Not surprisingly, the evaluation dataset shows the same results:

```
## [1] "type:" "mean"
## [1] "r2mean:" "0.4031"
```

```
## [1] "r2median:" "0.403"
## [1] "r2omit" "0.4019"
```

Although outliers and possible bad data appear in a number of places, without domain knowledge we are reluctant to eliminate any other outliers or influential points at this point without good reason. We don't know if extreme numbers are necessarily implausible. Therefore the outliers will remain.

3. Data Transformation

1. We create a flag for hits under 1500

As previously noted, Pitching_H is surprisingly weak in it's relationship to wins, and in fact appears positive for a large portion of its distribution. We examine more closely the relationship between pitching hits and wins, paying particular attention to the portion of the relationship where hits are below 3,000 (fig. 7).

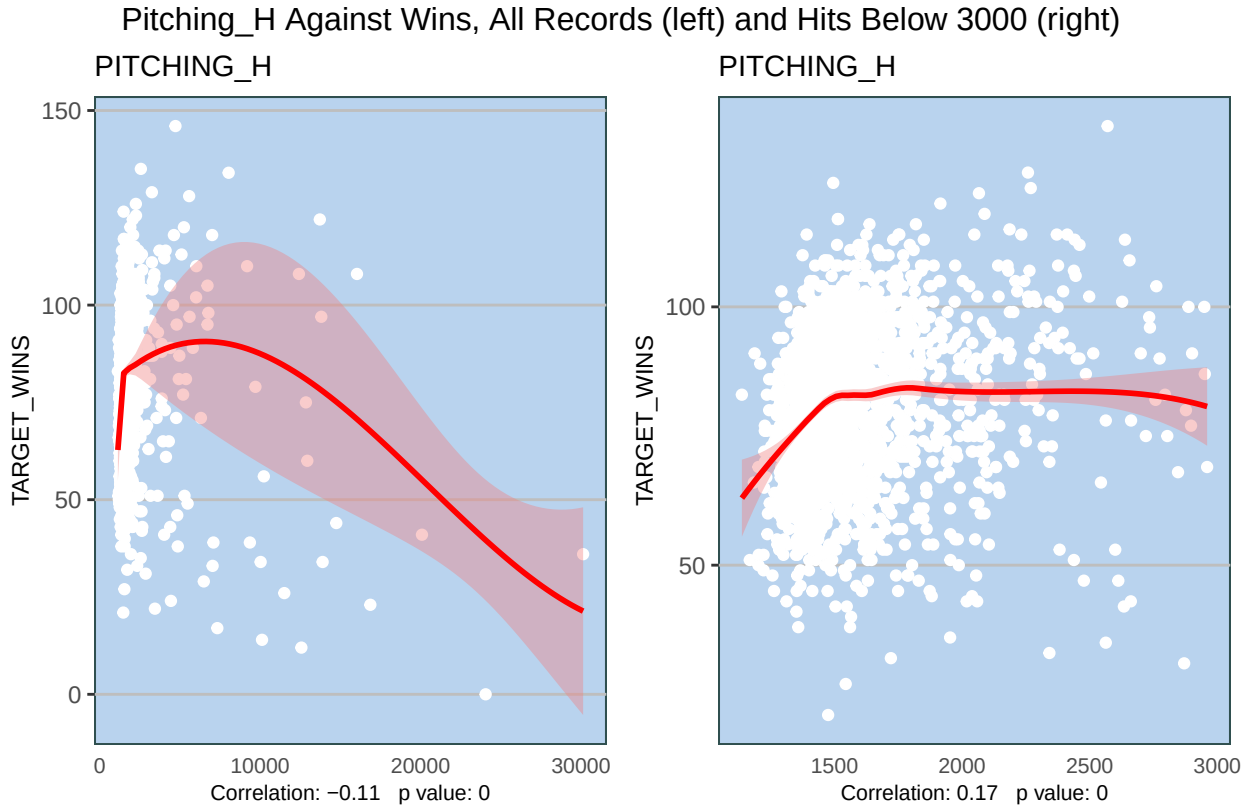


Fig.7

We can see here the positive correlation between pitching_h and wins. While we can't explain the phenomenon, we can account for it statistically by adding a binary flag for records with hits under 1500.

2. We create an interaction between Fielding_DP and hits.

The Fielding_DP correlation with Target Wins is surprising, since making double plays should help a team win. On the other hand, a team that makes double plays is also a team that gives up hits.

We therefore create an interaction term for Fielding_DP and Pitching_H.

3. We drop PITCHING_HR because it is an implausibly close match with HITTING_HR.

Like many pitching columns, Pitching_HR is unexpectedly positively correlated with wins. However, what makes this column truly implausible is how close a match it is with BATTING_HR. The scatterplot below

(Fig. 8) shows that the vast majority of the figures for pitching HR are exactly the same or within 2 or 3 of Batting HR. We therefore drop it since this makes no sense.

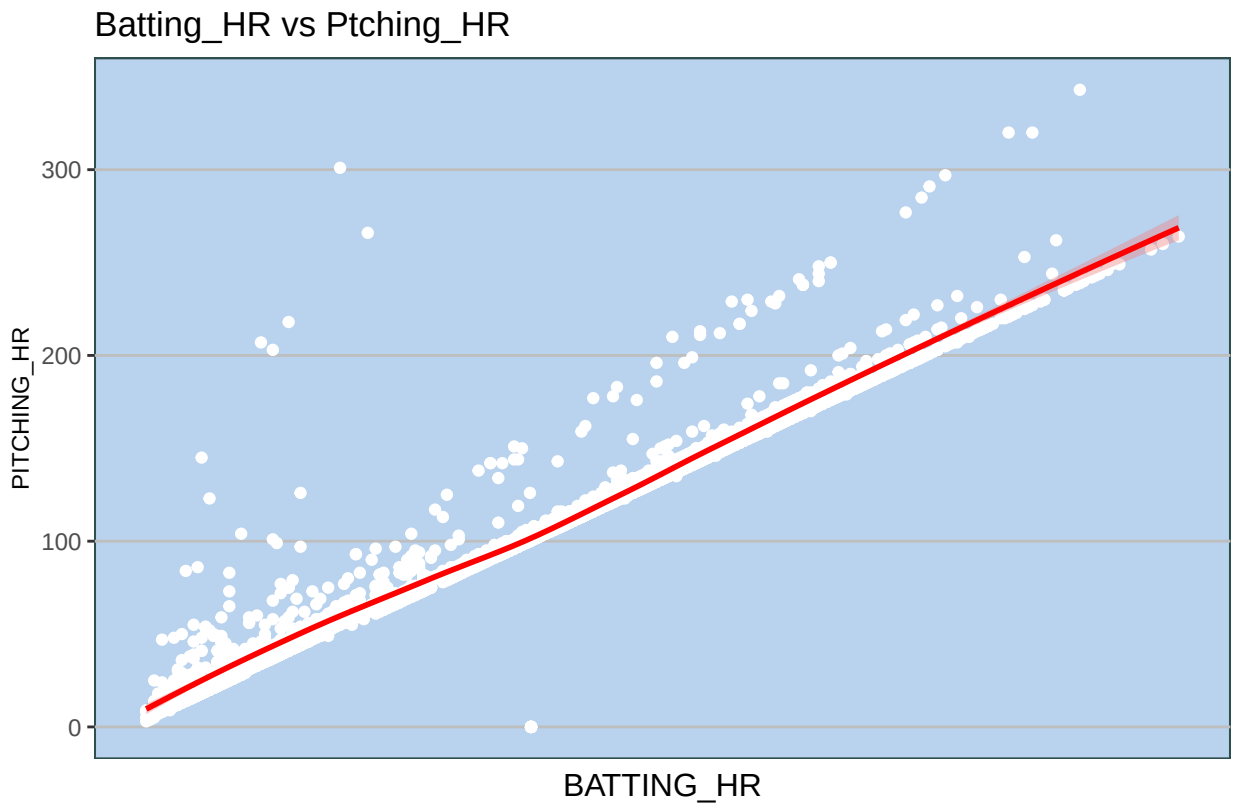


Fig. 8

4. We create a flag to account for the bimodal distribution of Batting HR.

Batting HR has a bimodal distribution (see Fig. 9). We don't explain this, but speculate that it may be related to different eras of baseball. Therefore, we create a flag to separate records with less than 80 HR form those with more.

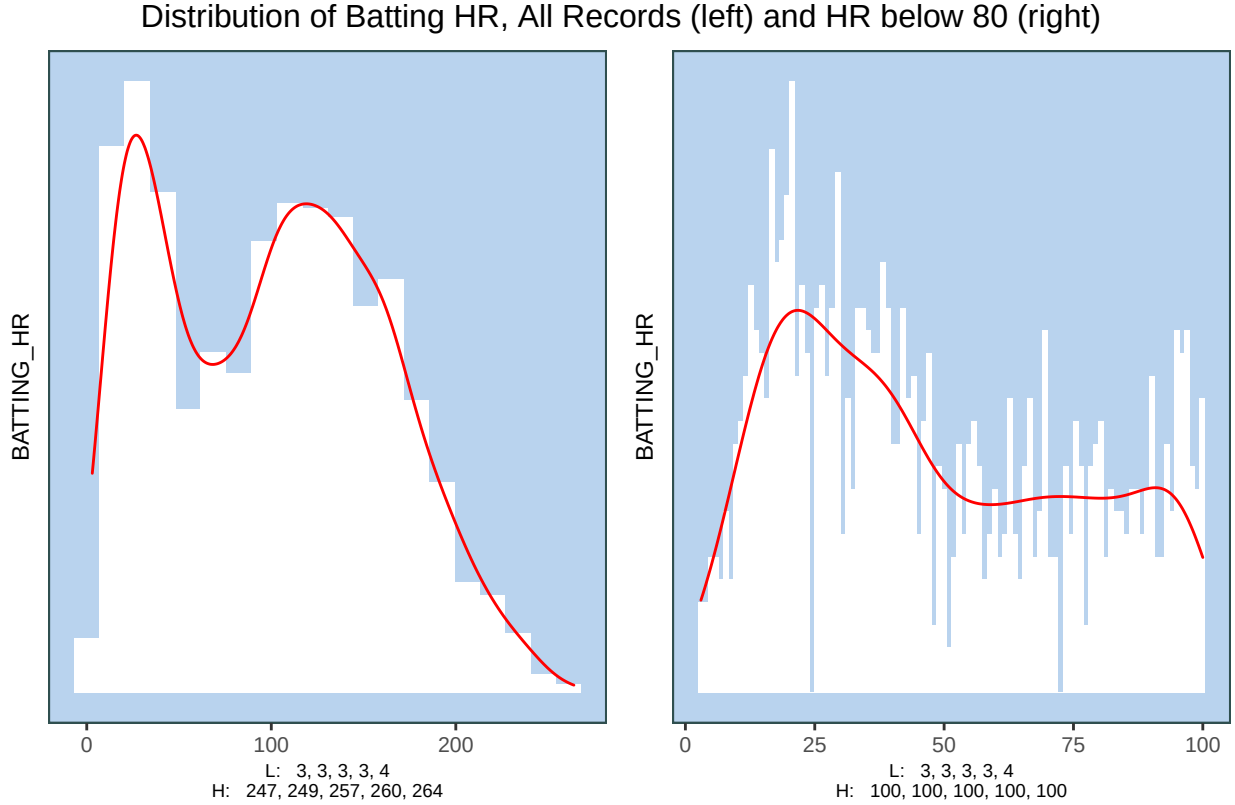


Fig.9

The numbers below represents the r-squareds of a simple linear regression of the column in question on the target variable before and after the transformation:

```
## [1] 0.02231354
```

```
## [1] 0.03503598
```

5. We transform the Fielding_err variable.

While the distributions of a number of columns suggest possible tranformations, we focus here on fielding errors, which has an upside-down u shape when correlated with wins. We therefore add an error squared term to the dataset.

The numbers below represents the r-squareds of a simple linear regression of the column in question on the target variable before and after the transformation:

```
## [1] 0.03072081
```

```
## [1] 0.04825783
```

6. We create interaction terms between the SO missing cohort and the columns identified above in the interaction analysis - Pitching_BB, Fielding E, Batting_H, Batting_HR, Batting_BB, Baserun_SB

The new fields are: Interaction_pbb_With_SO_Missing, Interaction_err_With_SO_Missing, Interaction_bh_With_SO_Missing, Interaction_bhr_With_SO_Missing, and Interaction_bbb_With_SO_Missing, Interaction_sb_With_SO_Missing.

7. For the sake of legibility, we do not create log terms for the many skewed distributions.

We would normally sacrifice some legibility for improved predictability by trying some log transformations on skewed independent variable distributions. However, legibility is already in serious peril with the odd behavior of the many pitching terms which suggest bad defense wins games. We therefore leave our transformations at those described.

4. Data Modeling

Here we build and test our models to gain insight into the dataset and ultimately predict outcomes.

According to the assignmrnt: "Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.)." As I have learned automated and manual selection in another class, I will use automated selection, in particular the "stepAIC" package.

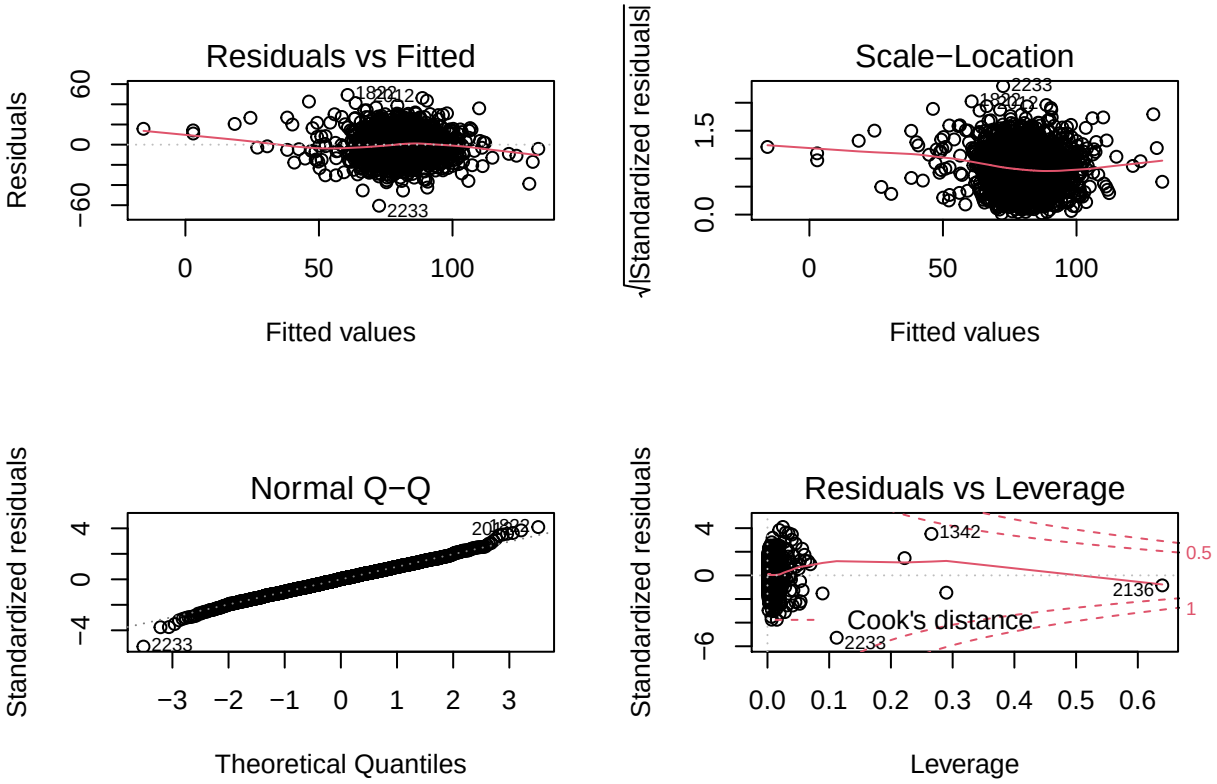
The stepAIC() function performs backward model selection by starting from a "maximal" model, which is then trimmed down. As each variable is eliminated, the Akaike Information Criterion (AIC) is calculated." The process stops when the AIC cannot be reduced by the elimnation of variables.

Because we are interested in interpretation as well as prediciton, we will modify the StepAIC model if we believe it improves readability.

a. Regression 1: Baseline (No transformations except flags for missing data)

```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B +
##     BATTING_HR + BATTING_BB + BATTING_SO + BASERUN_SB + PITCHING_H +
##     PITCHING_SO + FIELDING_E + FIELDING_DP + BSO_Missing_Flag +
##     BRSB_Missing_Flag + FDP_Missing_Flag, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.531  -8.063   0.330   8.075  49.266
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       13.7948052  5.0143117   2.751  0.00599 **
## BATTING_H          0.0521109  0.0033520  15.546  < 2e-16 ***
## BATTING_2B        -0.0401259  0.0086621  -4.632 3.82e-06 ***
## BATTING_3B         0.0537762  0.0158617   3.390  0.00071 ***
## BATTING_HR         0.0595856  0.0089648   6.647 3.75e-11 ***
## BATTING_BB         0.0260490  0.0032618   7.986 2.20e-15 ***
## BATTING_SO        -0.0066440  0.0022278  -2.982  0.00289 **
## BASERUN_SB         0.0477764  0.0046194  10.343  < 2e-16 ***
## PITCHING_H         0.0018926  0.0003398   5.569 2.86e-08 ***
## PITCHING_SO       -0.0013966  0.0006654  -2.099  0.03593 *
## FIELDING_E        -0.0560670  0.0033748 -16.613  < 2e-16 ***
## FIELDING_DP       -0.0969459  0.0134629  -7.201 8.10e-13 ***
## BSO_Missing_Flag   8.3474206  1.4721894   5.670 1.61e-08 ***
## BRSB_Missing_Flag 34.1064444  1.8484454  18.451  < 2e-16 ***
## FDP_Missing_Flag   4.2303099  1.4669785   2.884  0.00397 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 12.17 on 2261 degrees of freedom
## Multiple R-squared:  0.4068, Adjusted R-squared:  0.4031
## F-statistic: 110.7 on 14 and 2261 DF,  p-value: < 2.2e-16
##
## [1] "VIF Analysis"
##         BATTING_H        BATTING_2B        BATTING_3B        BATTING_HR
##          3.608349          2.524443          3.016545          4.444255
##        BATTING_BB        BATTING_SO        BASERUN_SB        PITCHING_H
##          2.459248          4.131567          2.380761          3.511045
##       PITCHING_SO        FIELDING_E       FIELDING_DP  BSO_Missing_Flag
##          1.946738          9.076248          1.674220          1.425731
## BRSB_Missing_Flag  FDP_Missing_Flag
##          2.848146          3.633432
```




```
## NULL
##
##  studentized Breusch-Pagan test
##
## data:  step3
## BP = 306.77, df = 14, p-value < 2.2e-16
##
##
##  Shapiro-Wilk normality test
##
## data:  step3$residuals
```

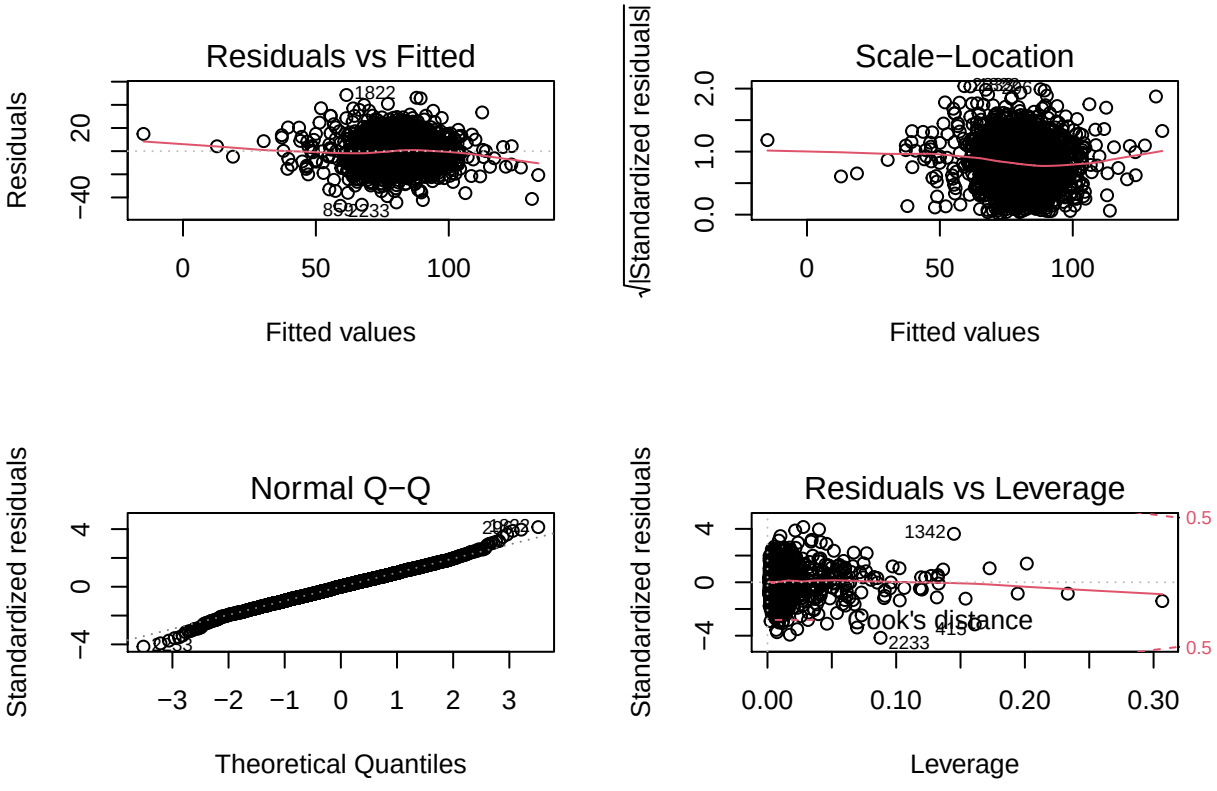
```
## W = 0.99701, p-value = 0.0002005
```

The adjusted r squared is .403. As we expected, many of the signs are in the "wrong" direction, especially for pitching. Without understanding why, we risk proceeding with a faulty model.

b. Regression 2: Include All transformations

```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B +
##     BATTING_HR + BATTING_BB + BATTING_SO + BASERUN_SB + PITCHING_H +
##     FIELDING_E + FIELDING_DP + BSO_Missing_Flag + BRSB_Missing_Flag +
##     FDP_Missing_Flag + Pitch_h_Under1500 + DP_times_PH + Fielding_Errors_sq +
##     Interaction_pbb_With_SO_Missing + Interaction_err_With_SO_Missing +
##     Interaction_bhr_With_SO_Missing + Interaction_bbb_With_SO_Missing +
##     Interaction_sb_With_SO_Missing, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.202  -7.806   0.193   7.821  48.504
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.472e+01  6.702e+00   3.688 0.000231 ***
## BATTING_H                       5.622e-02  3.302e-03  17.023  < 2e-16 ***
## BATTING_2B                     -4.125e-02  8.586e-03  -4.805 1.65e-06 ***
## BATTING_3B                      6.743e-02  1.610e-02   4.188 2.93e-05 ***
## BATTING_HR                      5.825e-02  8.978e-03   6.488 1.06e-10 ***
## BATTING_BB                      2.593e-02  3.247e-03   7.984 2.23e-15 ***
## BATTING_SO                     -1.223e-02  2.218e-03  -5.512 3.95e-08 ***
## BASERUN_SB                      5.238e-02  4.795e-03  10.923  < 2e-16 ***
## PITCHING_H                     -4.629e-03  2.995e-03  -1.546 0.122287
## FIELDING_E                     -8.282e-02  7.453e-03 -11.112  < 2e-16 ***
## FIELDING_DP                    -1.646e-01  3.571e-02  -4.610 4.25e-06 ***
## BSO_Missing_Flag                5.042e+01  1.190e+01   4.237 2.36e-05 ***
## BRSB_Missing_Flag               3.794e+01  2.023e+00  18.752  < 2e-16 ***
## FDP_Missing_Flag                5.282e+00  1.713e+00   3.084 0.002064 **
## Pitch_h_Under1500               2.214e+00  6.829e-01   3.242 0.001206 **
## DP_times_PH                     3.671e-05  2.040e-05   1.799 0.072094 .
## Fielding_Errors_sq              2.143e-05  4.284e-06   5.002 6.11e-07 ***
## Interaction_pbb_With_SO_Missing  1.336e-01  8.560e-02   1.560 0.118847
## Interaction_err_With_SO_Missing -1.938e-01  2.809e-02  -6.899 6.77e-12 ***
## Interaction_bhr_With_SO_Missing  3.652e-01  1.546e-01   2.362 0.018285 *
## Interaction_bbb_With_SO_Missing -1.397e-01  9.536e-02  -1.465 0.143105
## Interaction_sb_With_SO_Missing   3.896e-02  2.653e-02   1.469 0.142097
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.89 on 2254 degrees of freedom
## Multiple R-squared:  0.4357, Adjusted R-squared:  0.4305
## F-statistic: 82.88 on 21 and 2254 DF,  p-value: < 2.2e-16
##
## [1] "VIF Analysis"
```

```
##                       BATTING_H                       BATTING_2B
##                        3.670470                         2.599487
##                       BATTING_3B                       BATTING_HR
##                        3.258269                         4.671215
##                       BATTING_BB                       BATTING_SO
##                        2.554246                         4.292413
##                      BASERUN_SB                       PITCHING_H
##                        2.688323                       285.743188
##                      FIELDING_E                      FIELDING_DP
##                       46.392975                        12.345591
##                 BSO_Missing_Flag                 BRSB_Missing_Flag
##                       97.619815                         3.575494
##                 FDP_Missing_Flag                 Pitch_h_Under1500
##                        5.189563                         1.863892
##                     DP_times_PH                Fielding_Errors_sq
##                       282.770320                        24.339858
## Interaction_pbb_With_SO_Missing Interaction_err_With_SO_Missing
##                      1095.501873                        50.949700
## Interaction_bhr_With_SO_Missing Interaction_bbb_With_SO_Missing
##                        7.645153                      1173.699962
##  Interaction_sb_With_SO_Missing
##                       21.438192
```




```
## NULL
##
##  studentized Breusch-Pagan test
```

```
##
## data:  step3
## BP = 348.01, df = 21, p-value < 2.2e-16
##
##
##  Shapiro-Wilk normality test
##
## data:  step3$residuals
## W = 0.99694, p-value = 0.0001613
```

We note that the StepAIC process included a number of variables even though they were not significant.

The second model has an adjusted r squared of .4305. This is not much better (although an ANOVA shows the p value of the improvement to be near 0). However the interpretive value of the model is greatly increased, as the coefficient signs are much more reasonable (except for Batting_2nd, which we will disregard here.)

We examine the residual plots in the model selection phase.

```
## Analysis of Variance Table
##
## Model 1: TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##     BATTING_BB + BATTING_SO + BASERUN_SB + PITCHING_H + PITCHING_SO +
##     FIELDING_E + FIELDING_DP + BSO_Missing_Flag + BRSB_Missing_Flag +
##     FDP_Missing_Flag
## Model 2: TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##     BATTING_BB + BATTING_SO + BASERUN_SB + PITCHING_H + FIELDING_E +
##     FIELDING_DP + BSO_Missing_Flag + BRSB_Missing_Flag + FDP_Missing_Flag +
##     Pitch_h_Under1500 + DP_times_PH + Fielding_Errors_sq + Interaction_pbb_With_SO_Missing +
##     Interaction_err_With_SO_Missing + Interaction_bhr_With_SO_Missing +
##     Interaction_bbb_With_SO_Missing + Interaction_sb_With_SO_Missing
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   2261 334871
## 2   2254 318536  7     16335 16.513 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. Regression 3: Aggregated Power Stats by Hitting and Pitching There are many more transformations possible, but we are interested here in trying a different direction - simplifying as opposed to creating a more complex model.

Throughout the analysis we have been struggling with a multicollinearity issue which we might characterize as follows:

___* Teams have limited budgets. Therefore, those with good batting may have weak pitching and vice-versa. Of course good pitching and good hitting win games - but for an individual team, the question is which wins more games - good hitting or good batting.*___

We begin by creating simple Power Hitting and Pitching Weakness scores for each team. We do this by applying a score of 1 to 5 (1 = 20th percentile and below, 5 = 80th percentile and above) for the Batting and Pitching H and BB columns of each team compared to the overall distribution. We add the pitching scores together to get a Pitching Weakness score and the batting scores for a Batting Strength score. We also subtract weakness from strength to get a Total Power score.

The number below represents the correlation between Batting Power and Pitching Weakness. We can see they are highly correlated, as we suspected. Teams are needing to balance Hitting and pitching given a limited budget:

```
## [1] 0.7257795
```

These boxplots show the relationships in each power/weakness category to overall wins. We can see the paradox at work here - the higher the pitching weakness, the higher the batting power, and the higher the wins (see Fig. 10)

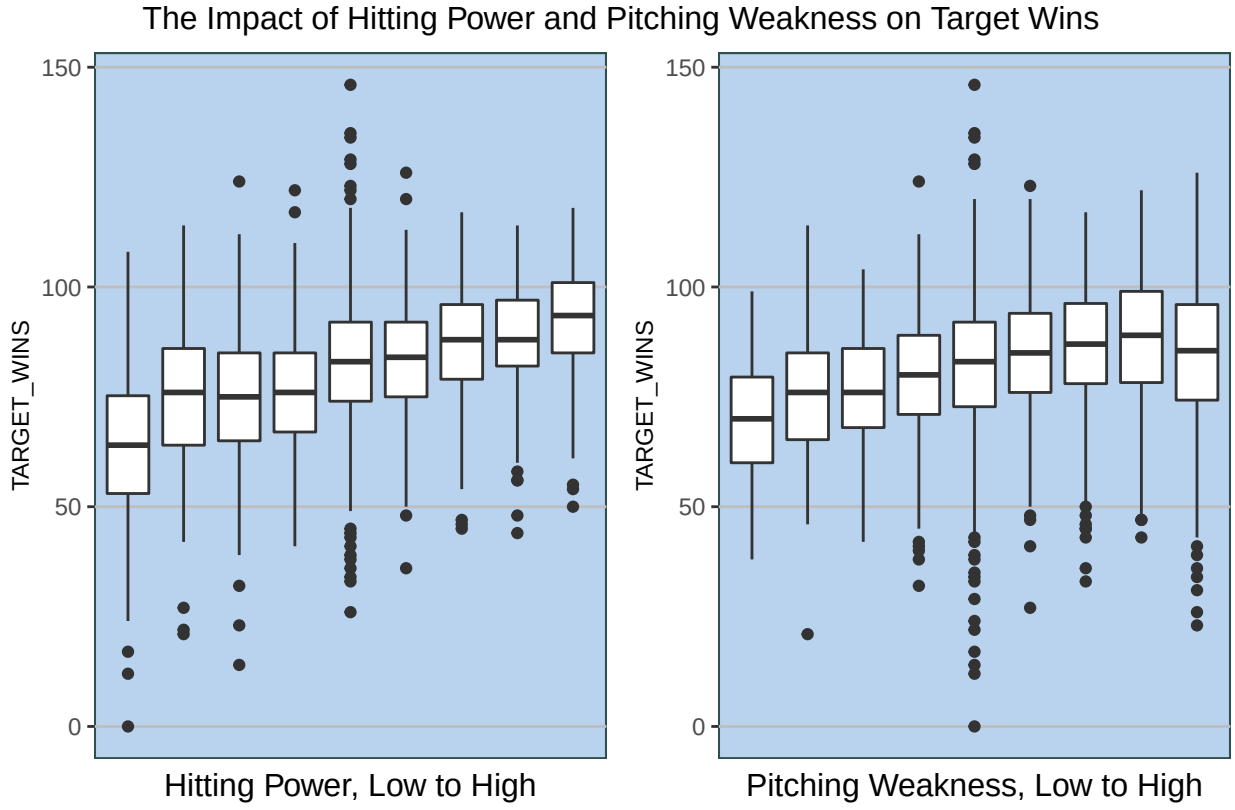


Fig. 10

We run regressions on total power, hitting power, pitching weakness and hitting power/pitching weakness combined.

```
##
## Call:
## lm(formula = TARGET_WINS ~ Total_Power, data = dfCat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.399  -9.788   0.518  10.365  65.212
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  80.7876     0.3247 248.844   <2e-16 ***
## Total_Power   1.8473     0.2076   8.898   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.49 on 2274 degrees of freedom
## Multiple R-squared:  0.03365,    Adjusted R-squared:  0.03322
## F-statistic: 79.18 on 1 and 2274 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ Hitting_Power, data = dfCat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.773  -8.841   0.201   9.159  65.159
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    62.7394     0.9260   67.75   <2e-16 ***
## Hitting_Power   3.0170     0.1462   20.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.46 on 2274 degrees of freedom
## Multiple R-squared:  0.1577, Adjusted R-squared:  0.1573
## F-statistic: 425.6 on 1 and 2274 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ Pitching_Weakness, data = dfCat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.825  -8.992   1.175  10.343  65.175
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        69.8306     0.9468   73.76   <2e-16 ***
## Pitching_Weakness   1.8323     0.1490   12.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.26 on 2274 degrees of freedom
## Multiple R-squared:  0.06238,    Adjusted R-squared:  0.06196
## F-statistic: 151.3 on 1 and 2274 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ Hitting_Power + Pitching_Weakness,
##     data = dfCat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.979  -8.829   0.436   9.162  65.162
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        63.6235     0.9732  65.373  < 2e-16 ***
## Hitting_Power       3.4647     0.2122  16.325  < 2e-16 ***
## Pitching_Weakness  -0.5957     0.2049  -2.907  0.00369 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 14.44 on 2273 degrees of freedom
## Multiple R-squared:  0.1608, Adjusted R-squared:   0.16
## F-statistic: 217.7 on 2 and 2273 DF,  p-value: < 2.2e-16
```

The model shows that in the balance between hitting and pitching, ***teams should emphasize good hitting and accept weak pitching.*** The adjusted r squared for the model with Hitting Power alone (.1573) is improved very little when pitching weakness is added to it (.16). The r-squared for Pitching Weakness alone is .06.

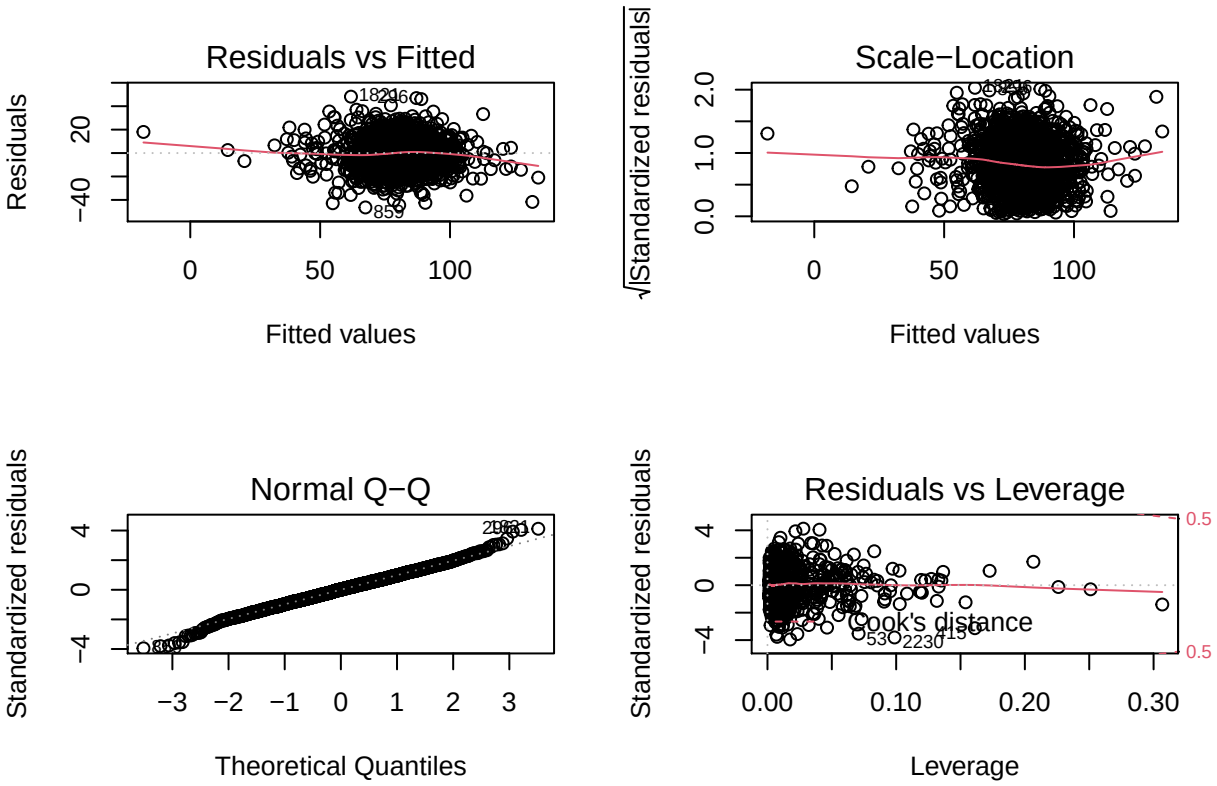
5. Model Selection

Now we select our model. The second model has the highest R squared and reliable interpretability so we will use it for our predictions. We will first eliminate the few influential points indicated by the residual plots.

We examine the new model's output:

```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B +
##     BATTING_HR + BATTING_BB + BATTING_SO + BASERUN_SB + PITCHING_H +
##     FIELDING_E + FIELDING_DP + BSO_Missing_Flag + BRSB_Missing_Flag +
##     FDP_Missing_Flag + Pitch_h_Under1500 + DP_times_PH + Fielding_Errors_sq +
##     Interaction_pbb_With_SO_Missing + Interaction_err_With_SO_Missing +
##     Interaction_bhr_With_SO_Missing + Interaction_bbb_With_SO_Missing +
##     Interaction_sb_With_SO_Missing, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.429  -7.819   0.225   7.856  48.134
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      2.374e+01  6.694e+00   3.546 0.000399 ***
## BATTING_H                        5.745e-02  3.312e-03  17.345  < 2e-16 ***
## BATTING_2B                      -4.386e-02  8.587e-03  -5.107 3.54e-07 ***
## BATTING_3B                       7.151e-02  1.610e-02   4.442 9.33e-06 ***
## BATTING_HR                       5.887e-02  8.958e-03   6.571 6.17e-11 ***
## BATTING_BB                       2.550e-02  3.241e-03   7.870 5.46e-15 ***
## BATTING_SO                      -1.222e-02  2.212e-03  -5.526 3.65e-08 ***
## BASERUN_SB                       5.202e-02  4.784e-03  10.872  < 2e-16 ***
## PITCHING_H                      -4.611e-03  2.996e-03  -1.539 0.123868
## FIELDING_E                      -8.444e-02  7.464e-03 -11.313  < 2e-16 ***
## FIELDING_DP                     -1.605e-01  3.568e-02  -4.499 7.16e-06 ***
## BSO_Missing_Flag                 5.031e+01  1.187e+01   4.240 2.33e-05 ***
## BRSB_Missing_Flag                3.728e+01  2.027e+00  18.393  < 2e-16 ***
## FDP_Missing_Flag                 5.210e+00  1.724e+00   3.023 0.002531 **
## Pitch_h_Under1500                2.243e+00  6.810e-01   3.294 0.001004 **
## DP_times_PH                      3.390e-05  2.042e-05   1.660 0.097107 .
## Fielding_Errors_sq               2.396e-05  4.335e-06   5.527 3.63e-08 ***
## Interaction_pbb_With_SO_Missing  1.347e-01  8.535e-02   1.578 0.114592
## Interaction_err_With_SO_Missing -1.939e-01  2.801e-02  -6.923 5.75e-12 ***
```

```
## Interaction_bhr_With_SO_Missing  3.605e-01  1.542e-01   2.338 0.019481 *
## Interaction_bbb_With_SO_Missing -1.403e-01  9.508e-02  -1.476 0.140113
## Interaction_sb_With_SO_Missing   3.857e-02  2.645e-02   1.458 0.144977
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.85 on 2251 degrees of freedom
## Multiple R-squared:  0.4384, Adjusted R-squared:  0.4331
## F-statistic: 83.67 on 21 and 2251 DF,  p-value: < 2.2e-16
##
## [1] "VIF Analysis"
##                        BATTING_H                        BATTING_2B
##                         3.707874                          2.607807
##                        BATTING_3B                        BATTING_HR
##                         3.260919                          4.673853
##                        BATTING_BB                        BATTING_SO
##                         2.547589                          4.287409
##                        BASERUN_SB                        PITCHING_H
##                         2.691815                        274.568838
##                        FIELDING_E                       FIELDING_DP
##                        46.562944                         12.396781
##                BSO_Missing_Flag                BRSB_Missing_Flag
##                        97.625954                          3.558545
##                FDP_Missing_Flag                Pitch_h_Under1500
##                         5.255244                          1.862688
##                       DP_times_PH              Fielding_Errors_sq
##                       271.957776                         25.000344
## Interaction_pbb_With_SO_Missing Interaction_err_With_SO_Missing
##                      1095.487226                         50.957782
## Interaction_bhr_With_SO_Missing Interaction_bbb_With_SO_Missing
##                         7.645237                       1173.677967
##  Interaction_sb_With_SO_Missing
##                        21.438268
```




```
## NULL
##
##  studentized Breusch-Pagan test
##
## data:  step3
## BP = 332.57, df = 21, p-value < 2.2e-16
##
##
##  Shapiro-Wilk normality test
##
## data:  step3$residuals
## W = 0.99713, p-value = 0.0003086
```

There is only a slight improvement with the elimination of influential points. We note that the adjusted r-squared is still best among all models. The F statistic shows the model is significant overall. The residual standard error is small relative to the target variable. We see no patterning in the residuals and the distribution is relatively normal, except at the tails.

There are new influential points after eliminating the others, but we accept them without any better reason to challenge them. A VIF analysis shows a fair amount of multicolinnearity, but we knew this, and even created more with our interaction terms. In all, we can move forward with this model without further modification.

6. Predictions

We use the model to make predictions. The entire predictions file is submitted separately in the appendices.

```
##   predict(m, newdata = dfEval2)
## 1                      65.55191
## 2                      73.00183
## 3                      74.25101
## 4                      69.95245
## 5                      66.80497
## 6                      70.24417
```

7. Conclusion

We examined ~2200 records of baseball teams to create a predictive model of wins. However, if this were an actual workplace project, it seems unlikely that the point would be the passive prediction of wins from sample performance statistics. Rather, the data would need to serve the question of what strategies should be employed to improve wins. Answering this question require more insight than ability to predict. Throughout this analysis we have confronted the counter-intuitive phenomenon that weaker pitching is correlated with better outcomes. ***Analysis shows that this is most likely because teams need to trade off pitching and hitting, and better hitting compensates more for poor pitching than vice versa. This is the most important finding of this examination.***