Predict 411

MONEYBALL ASSIGNMENT

Zeeshan Latifi

Northwestern University Winter 2018

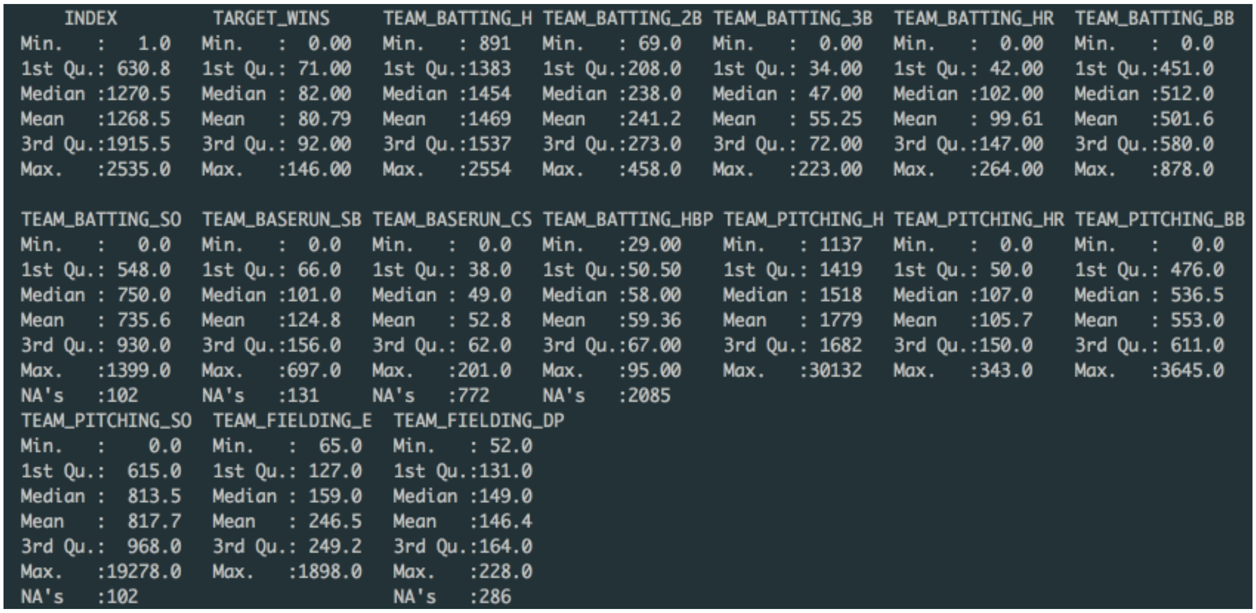
Introduction:

The data that is being used for OLS regression analysis contains approximately 2200 records. Each record represents 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. For this analysis we will conduct some initial exploratory data analysis, clean the data, impute missing values, build some models, and then compare each model. The goal for the model is to be an optimal predictor of games that a given team should win based on the data provided.

Data Exploration:

The data set contains 17 variables with 2276 observations. Figure 1 below highlights the summary of the data for each variable. As you can see there are several variables that contain missing values. The most being TEAM_BATTING_HBP, approximately 92% of this data is missing. Because of this, we will exclude this variable as part of our analysis. For the remaining variables, we will impute values for the those listed as 'NA'.

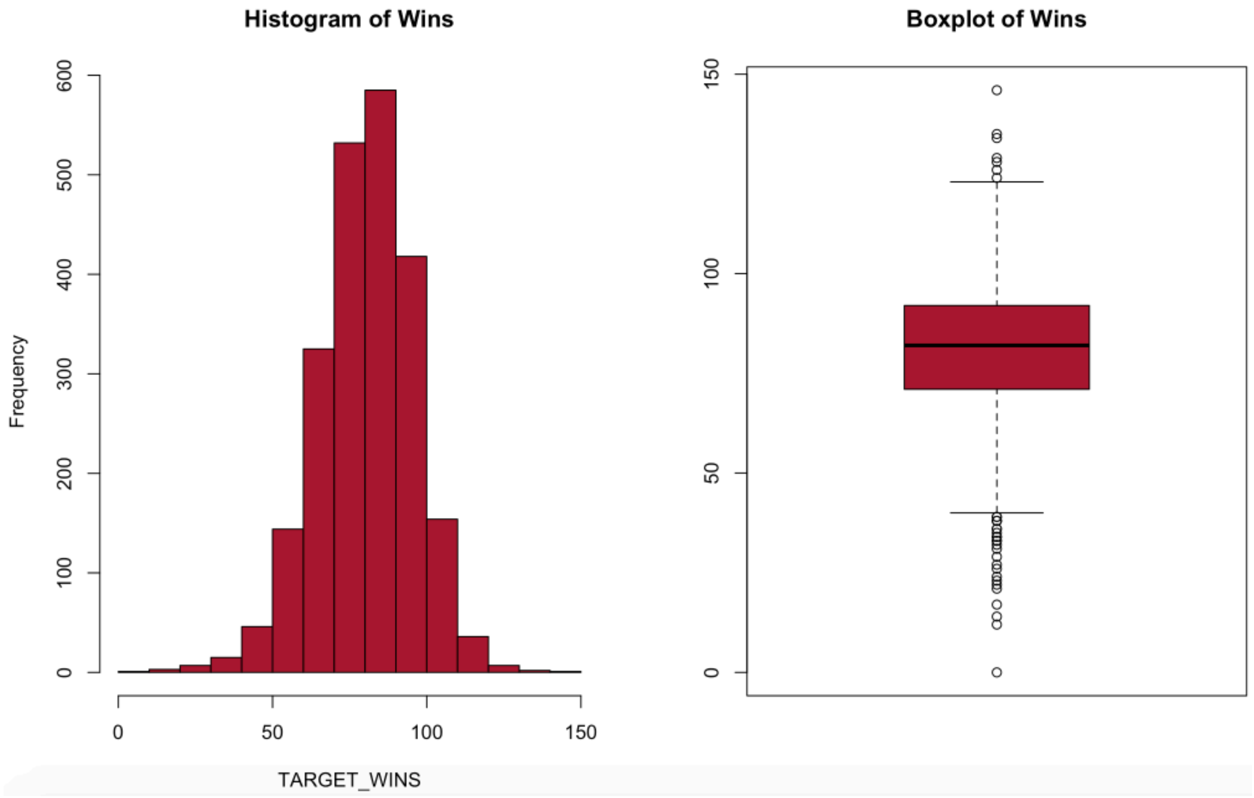
Figure 1



```
     INDEX         TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB
 Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0   Min.   :  0.00   Min.   :  0.00   Min.   :  0.0
 1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00   1st Qu.: 42.00   1st Qu.:451.0
 Median :1270.5   Median : 82.00   Median :1454   Median :238.0   Median : 47.00   Median :102.00   Median :512.0
 Mean   :1268.5   Mean   : 80.79   Mean   :1469   Mean   :241.2   Mean   : 55.25   Mean   : 99.61   Mean   :501.6
 3rd Qu.:1915.5   3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00   3rd Qu.:147.00   3rd Qu.:580.0
 Max.   :2535.0   Max.   :146.00   Max.   :2554   Max.   :458.0   Max.   :223.00   Max.   :264.00   Max.   :878.0

 TEAM_BATTING_SO  TEAM_BASERUN_SB TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
 Min.   :   0.0   Min.   :  0.0   Min.   :  0.0   Min.   :29.00    Min.   : 1137   Min.   :  0.0    Min.   :   0.0
 1st Qu.: 548.0   1st Qu.: 66.0   1st Qu.: 38.0   1st Qu.:50.50    1st Qu.: 1419   1st Qu.: 50.0    1st Qu.: 476.0
 Median : 750.0   Median :101.0   Median : 49.0   Median :58.00    Median : 1518   Median :107.0    Median : 536.5
 Mean   : 735.6   Mean   :124.8   Mean   : 52.8   Mean   :59.36    Mean   : 1779   Mean   :105.7    Mean   : 553.0
 3rd Qu.: 930.0   3rd Qu.:156.0   3rd Qu.: 62.0   3rd Qu.:67.00    3rd Qu.: 1682   3rd Qu.:150.0    3rd Qu.: 611.0
 Max.   :1399.0   Max.   :697.0   Max.   :201.0   Max.   :95.00    Max.   :30132   Max.   :343.0    Max.   :3645.0
 NA's   :102      NA's   :131     NA's   :772     NA's   :2085
 TEAM_PITCHING_SO  TEAM_FIELDING_E  TEAM_FIELDING_DP
 Min.   :    0.0   Min.   :  65.0   Min.   : 52.0
 1st Qu.:  615.0   1st Qu.: 127.0   1st Qu.:131.0
 Median :  813.5   Median : 159.0   Median :149.0
 Mean   :  817.7   Mean   : 246.5   Mean   :146.4
 3rd Qu.:  968.0   3rd Qu.: 249.2   3rd Qu.:164.0
 Max.   :19278.0   Max.   :1898.0   Max.   :228.0
 NA's   :102                        NA's   :286
```

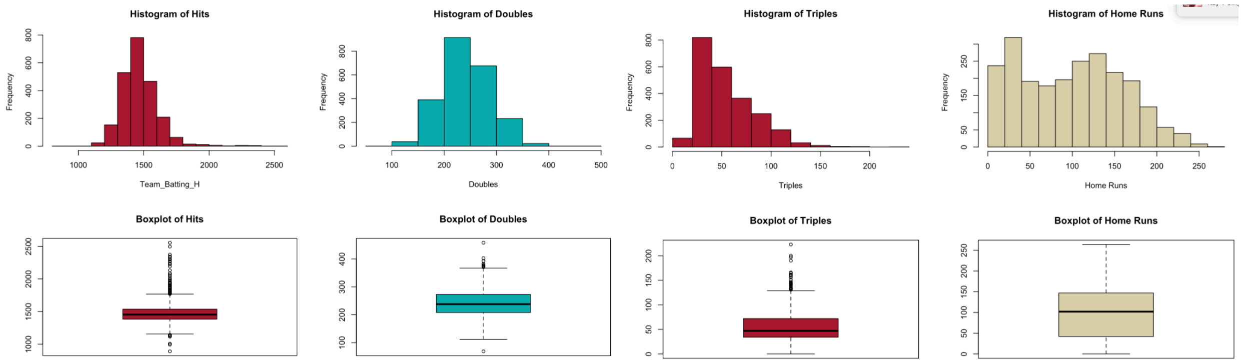
Next, we took a look at our response variable, TARGET_WINS. As shown in the figure below, there are several outliers. The data is somewhat normally distributed, with a slight skew to the right. Due to the somewhat normal distribution, the decision is the leave this variable as is. It may not be best to interfere with the values in the response variable for a best fit model.

Figure 2



Next we did some preliminary evaluations of the some of the predictor variables.

Figure 3



The data for hits and doubles are somewhat normally distributed with some outliers. However, when looking at the histogram for triples, you can see that it is heavily skewed to the right. This makes sense, there are not many games where many triples are hit. Home Runs is also a bit skewed, and not normally distributed; some transformations maybe required here.

Figure 4 below illustrates the distribution of the of Walks, Strikeouts, and Hits by Pitch. Walks are skewed left, also shown by the number of lower end outliers. Strikeouts do not have any outliers, however, the data could use some transformation to normalize. HBP, there is not a sufficient amount of data for use to utilize this as a reliable variable.

Figure 4

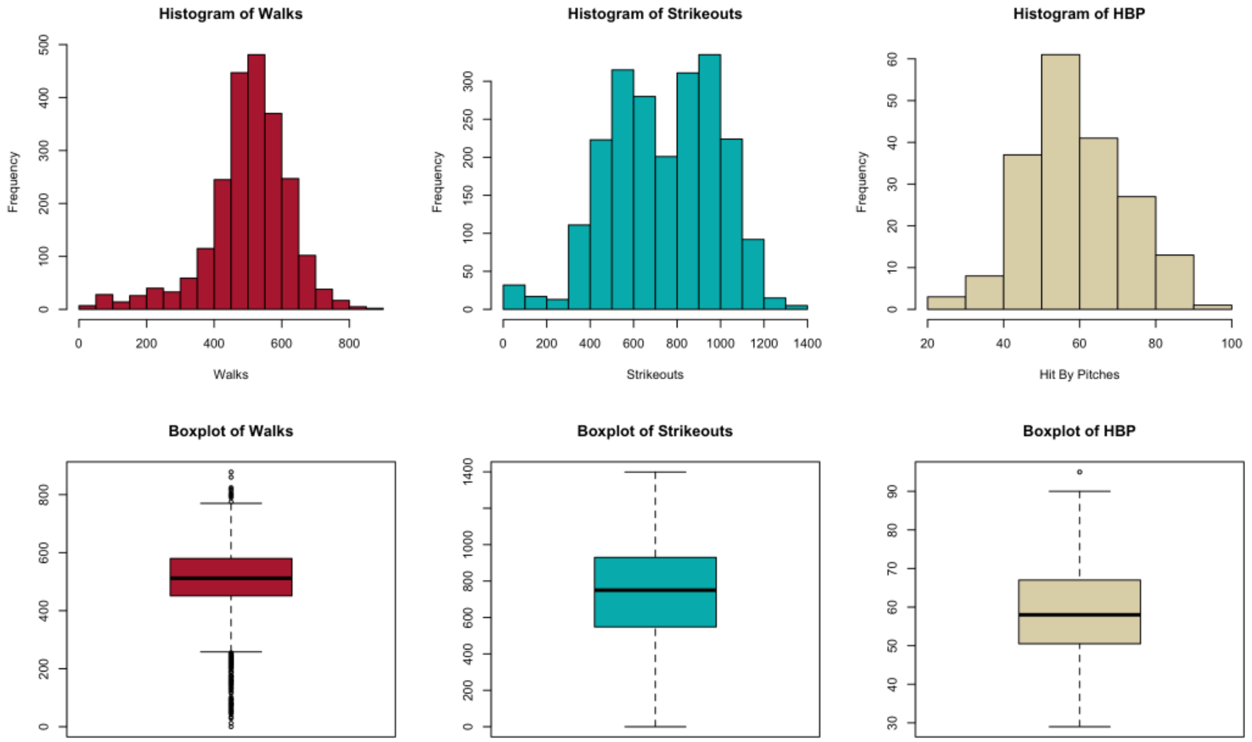
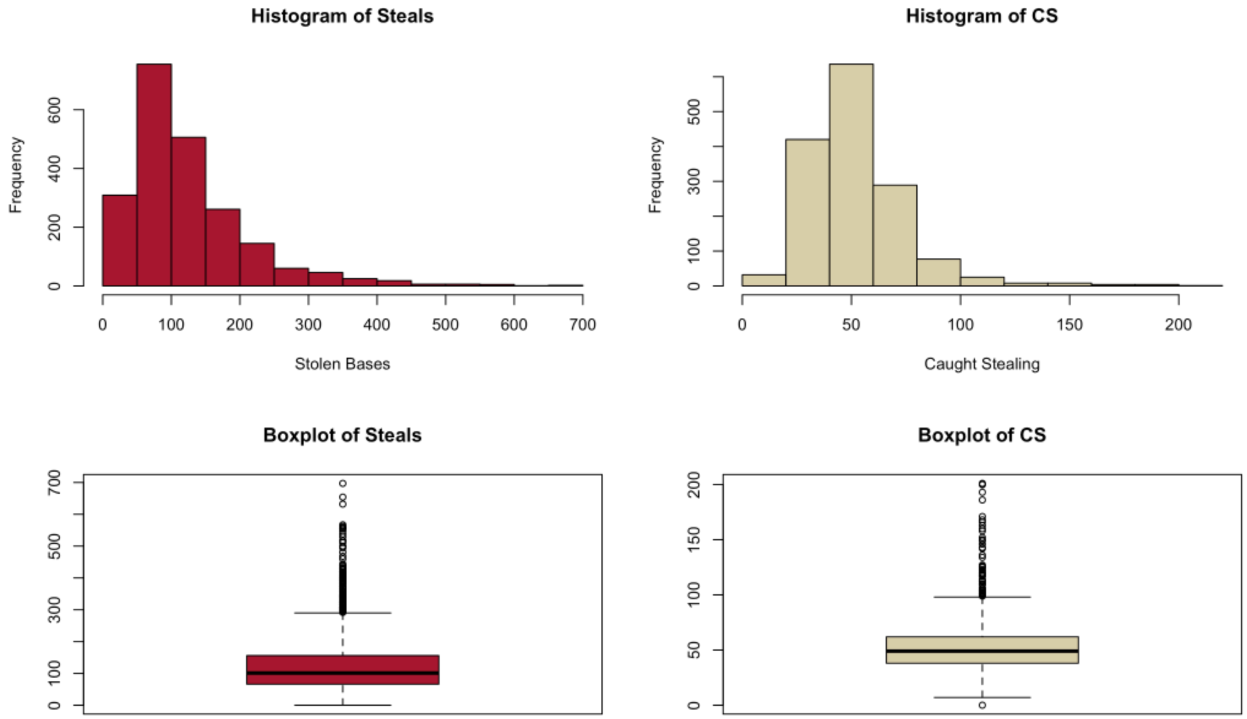


Figure 5 below details steals and caught stealing statistics. Both data are heavily skewed right also indicative of the outliers that are on the upper level. Some trimming may be required here when conducting our final models. Figure 6 below also shows heavy right skew for Hits Against. Home Runs Against is fairly distributed, however could benefit from a log transformation to normalize the data.

Figure 5



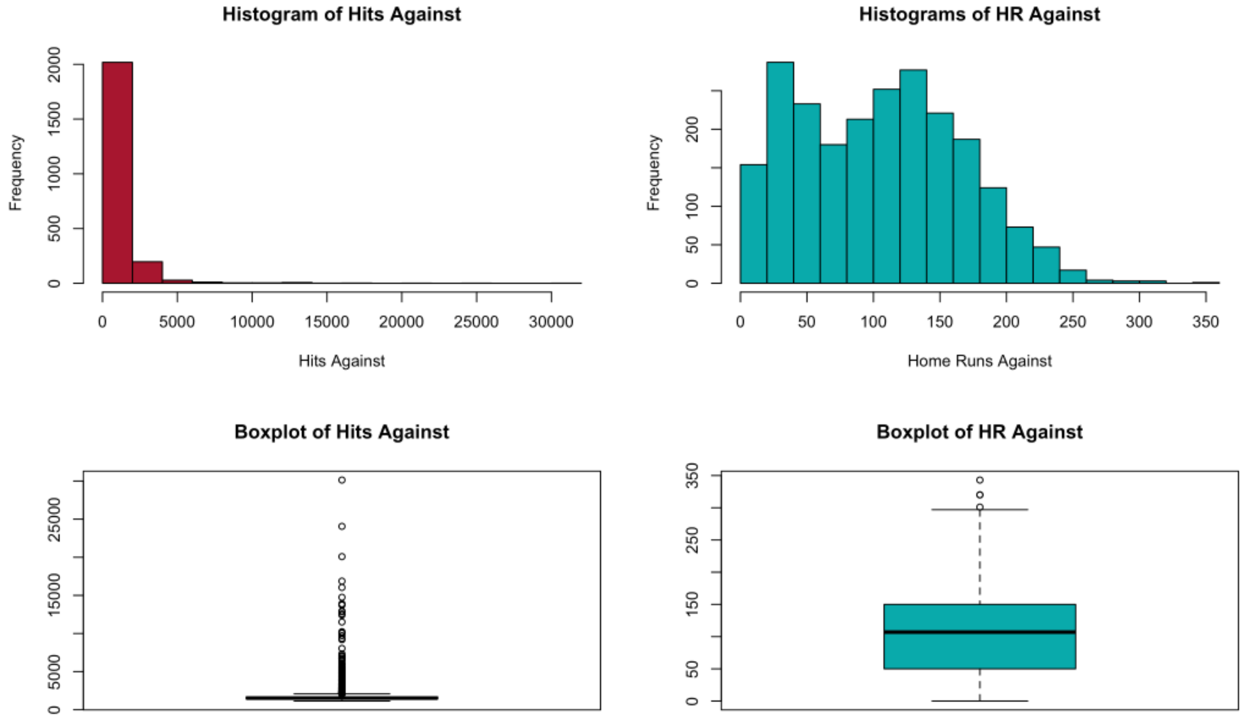
Histogram of Steals

Histogram of CS

Boxplot of Steals

Boxplot of CS

Figure 6



Histogram of Hits Against

Histograms of HR Against

Boxplot of Hits Against

Boxplot of HR Against

Figure 7

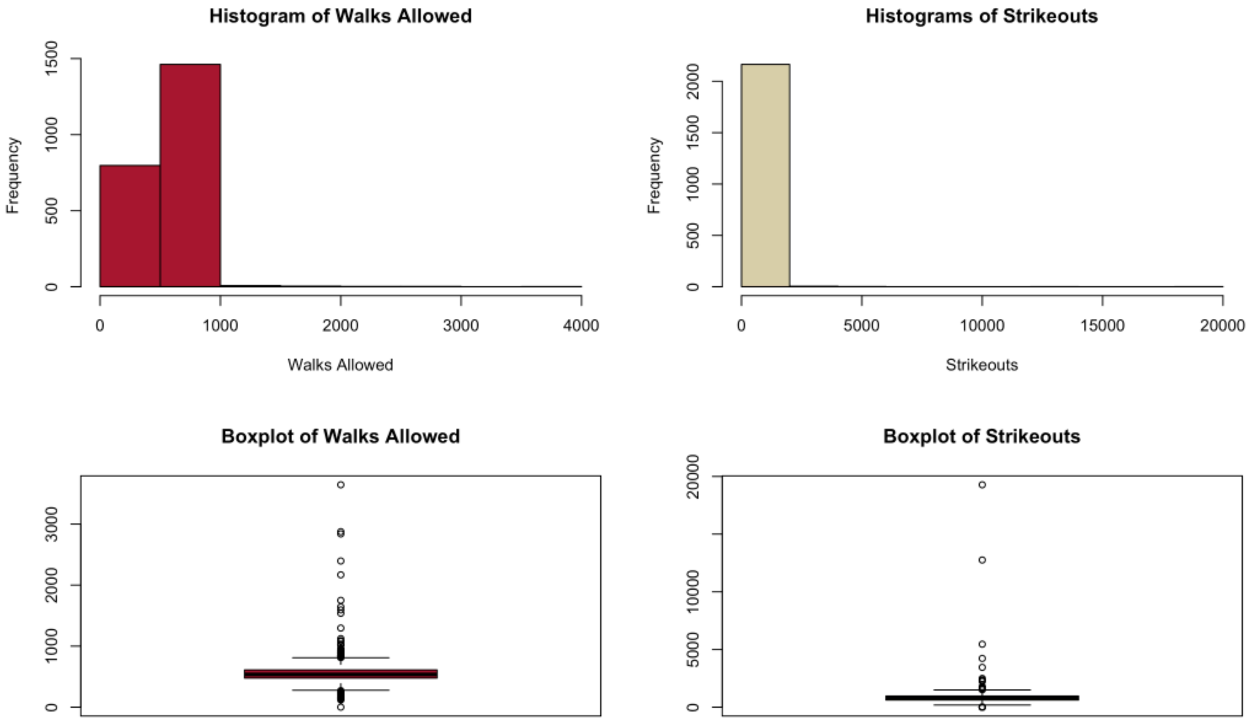
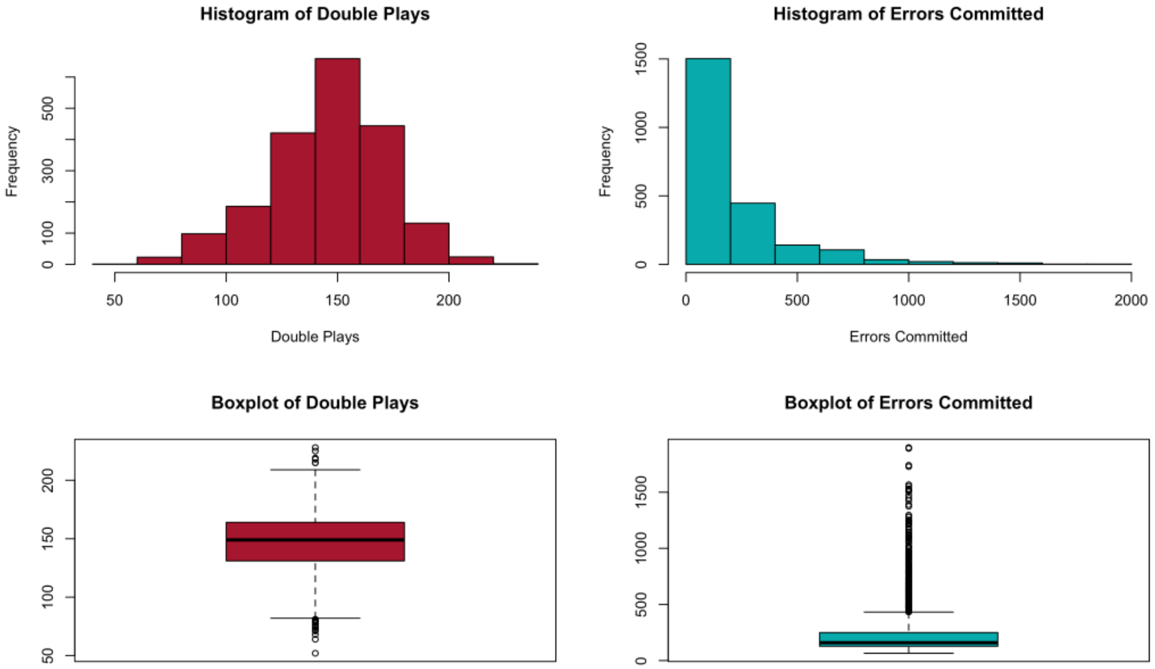


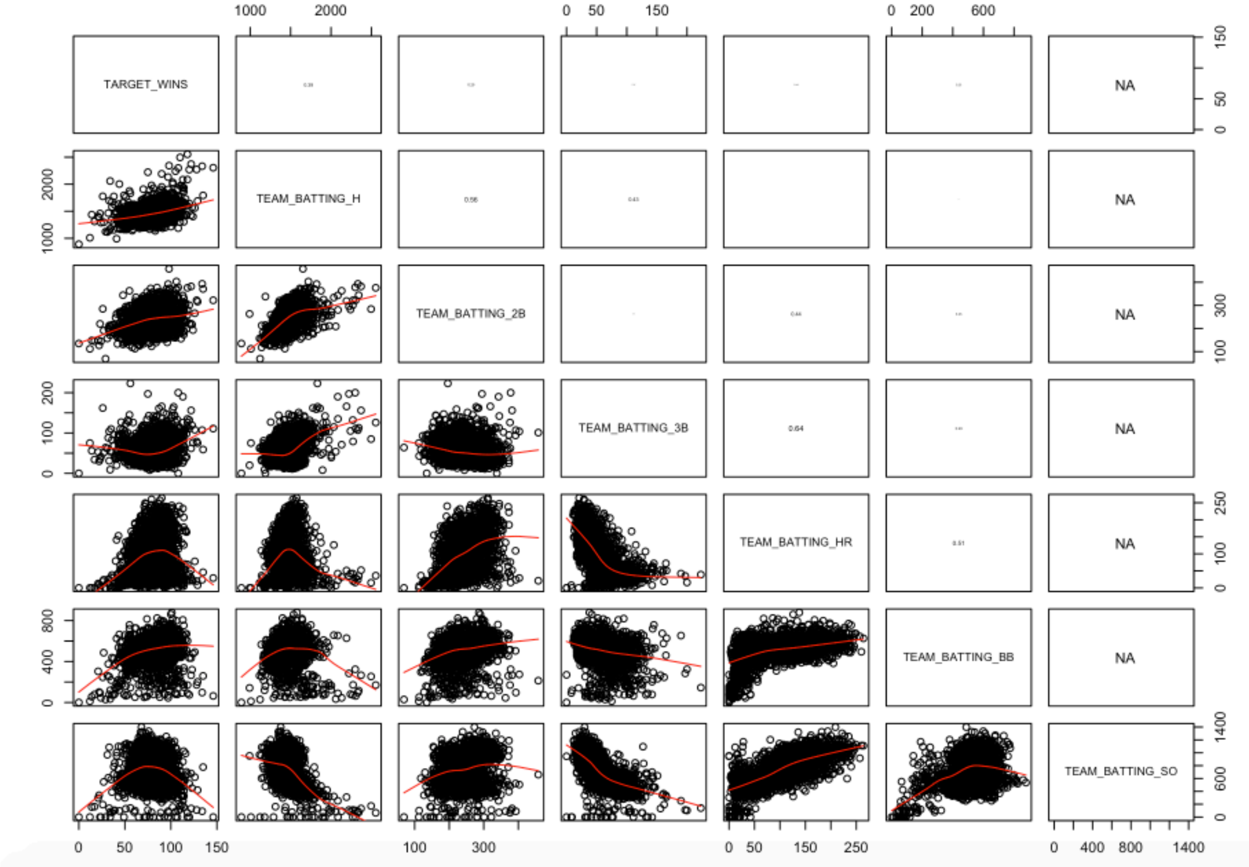
Figure 7, shown above has many outliers in Walks Allowed. However, some imputations will be required here for the missing values. Strikeouts are also skewed, with many missing values.

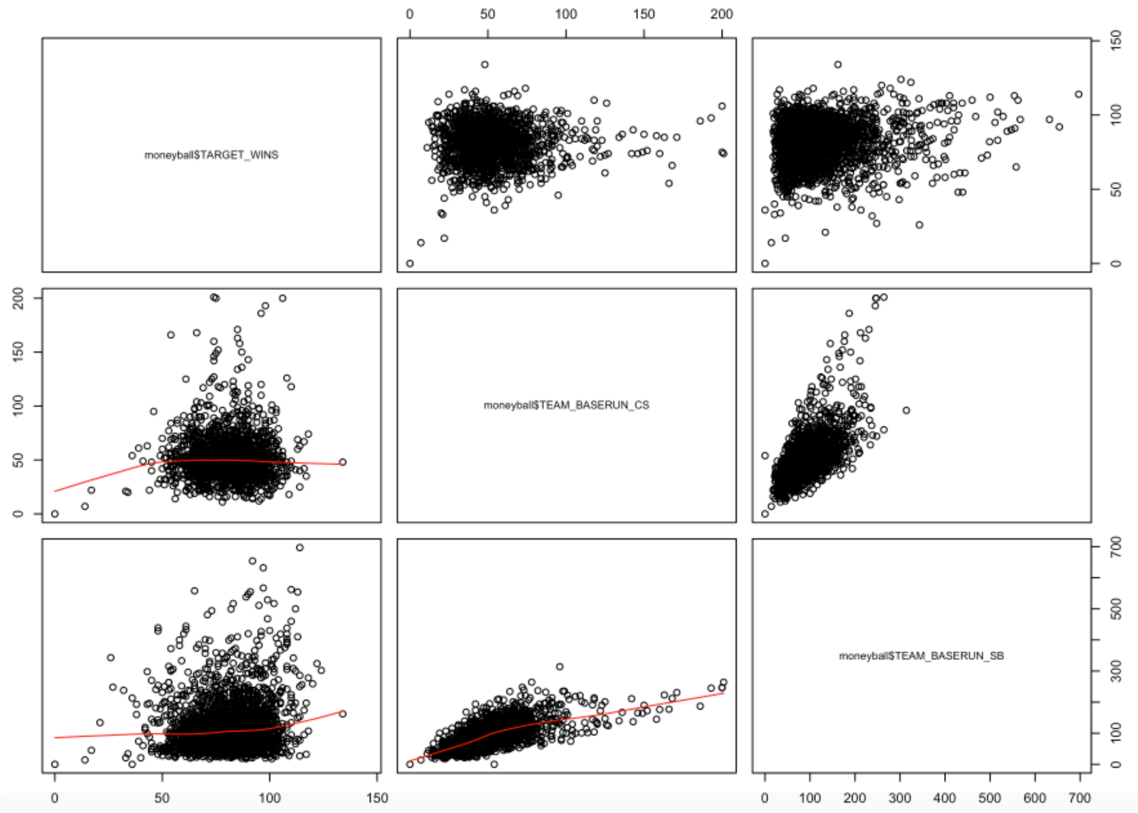
Figure 8



The Double Plays data in Figure 8 is somewhat normally distributed with a slight skew to the left. We will leave this as is. Errors committed, is heavily skewed, this makes sense due to baseball games generally don't have that many errors.

Figure 9




moneyball$TARGET_WINS
moneyball$TEAM_BASERUN_CS
moneyball$TEAM_BASERUN_SB
0
50
100
150
200
0
50
100
150
0
50
100
150
200
100
300
500
700
0
100
200
300
400
500
600
700

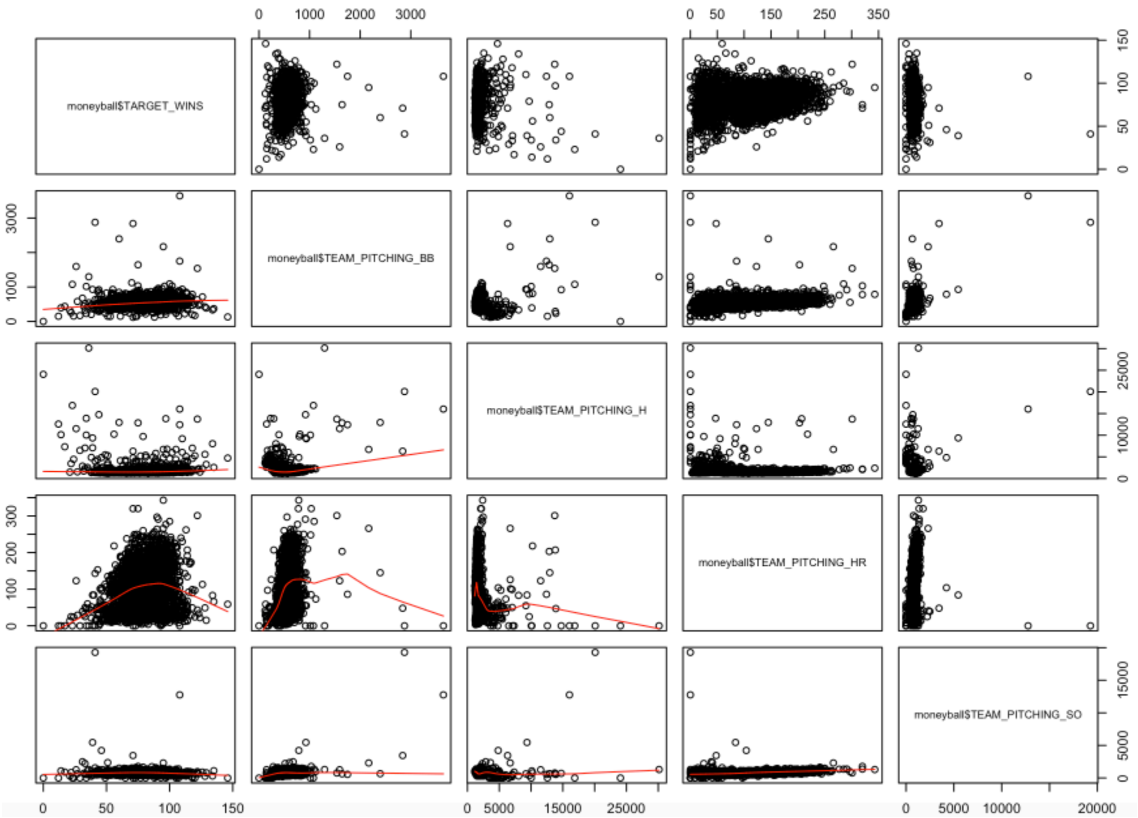
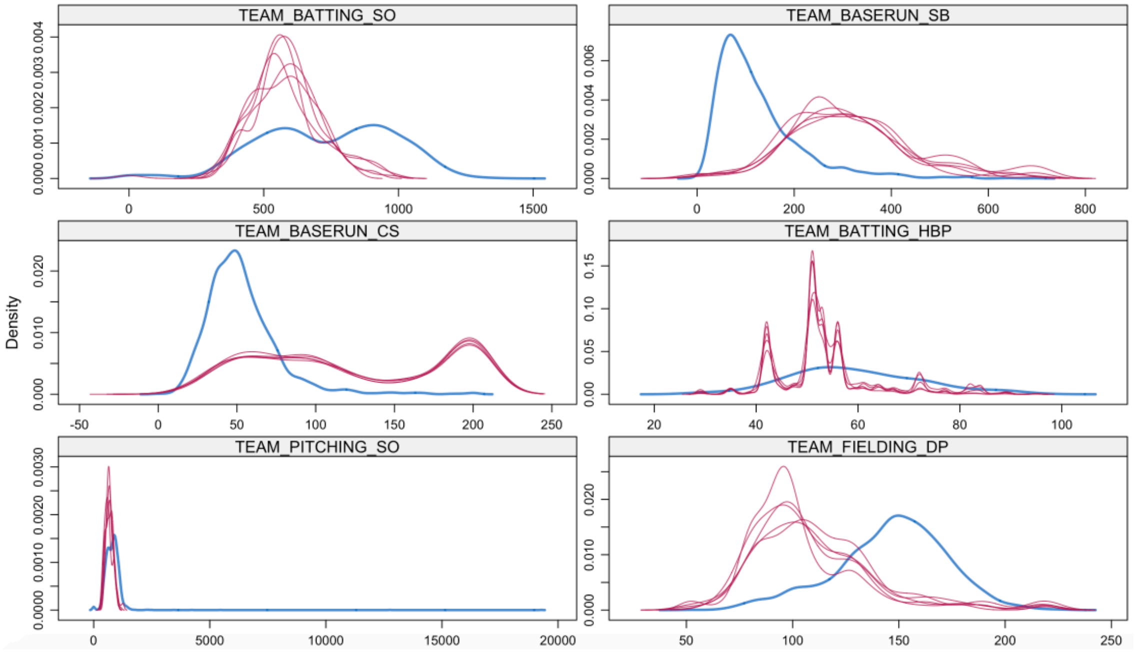

moneyball$TARGET_WINS
moneyball$TEAM_PITCHING_BB
moneyball$TEAM_PITCHING_H
moneyball$TEAM_PITCHING_HR
moneyball$TEAM_PITCHING_SO
0
1000
2000
3000
0
50
150
250
350
0
50
100
150
0
1000
3000
0
10000
25000
0
100
200
300
0
5000
15000
0
50
100
150
0
5000
15000
25000
0
5000
10000
20000

Figure 9 above has three components. Here we tested for correlations between the variables within our data set. In the first part, it can be seen that our response variable TARGET_WINS has a positive correlation with TEAM_BATTING_H and TEAM_BATTING_2B. This is of no surprise really, theoretically more hits should lead to more wins. However, further evaluation is needed. TARGET_WINS has little correlation when looking at the defensive statistics, as shown in the second part of Figure 9. However, we can see that TEAM_BASERUN_CS is positively correlated to TEAM_BASERUN_SB. This also theoretically makes sense; the more stolen bases you have the more likely it is that the player is going to get caught.

Data Preparation:

To start, we need to impute the missing values in our data. In order to this we've using the MICE package in R; specifically, the PMM (Predictive Mean Matching) method. To test the results using PMM, the below density plot was generated. As show, TEAM_BATTING_HBP is very much off. This is due to the lack of initial data that was provided. Because of this, we will forgo this variable. The other variables, are very similar to somewhat similar. These will be used in our assessment going forward.

Figure 10



After verifying that there were no further NA's in our data, we then moved onto transformations of certain variables and the creation on new variables. Below is the list of new variables:

- Singles hit by the batting team: TEAM_BATTING_1B = TEAM_BATTING_H - TEAM_BATTING_HR - TEAM_BATTING_3B - TEAM_BATTING_2B
- Team fielding error over 500, will be replaced with a value of 500

Log Transformations: All were used to normalize the data (per our findings during EDA)

- TEAM_BATTING_1B
- TEAM_BATTING_3B
- TEAM_BASERUN_SB
- TEAM_BASERUN_CS
- TEAM_BATTING_HR
- TEAM_BATTING_SO
- TEAM_PITCHING_HR

Next, we reevaluated the transformed data to conduct any trimming of the data. Based on our findings, we decided to trim:

- 21=< TARGET_WINS <= 120
- TEAM_PITCHING_H < 2000

Build Linear Regression Models:

From our cleansed data, we now begin building an optimal model for predicting our response variable: target wins. Variable selection came next. To do this, we used the regsubset() function in R and obtained a score for each variable. Based on the output in R, we were able to determine that the best model could be built using the following predictor variables:

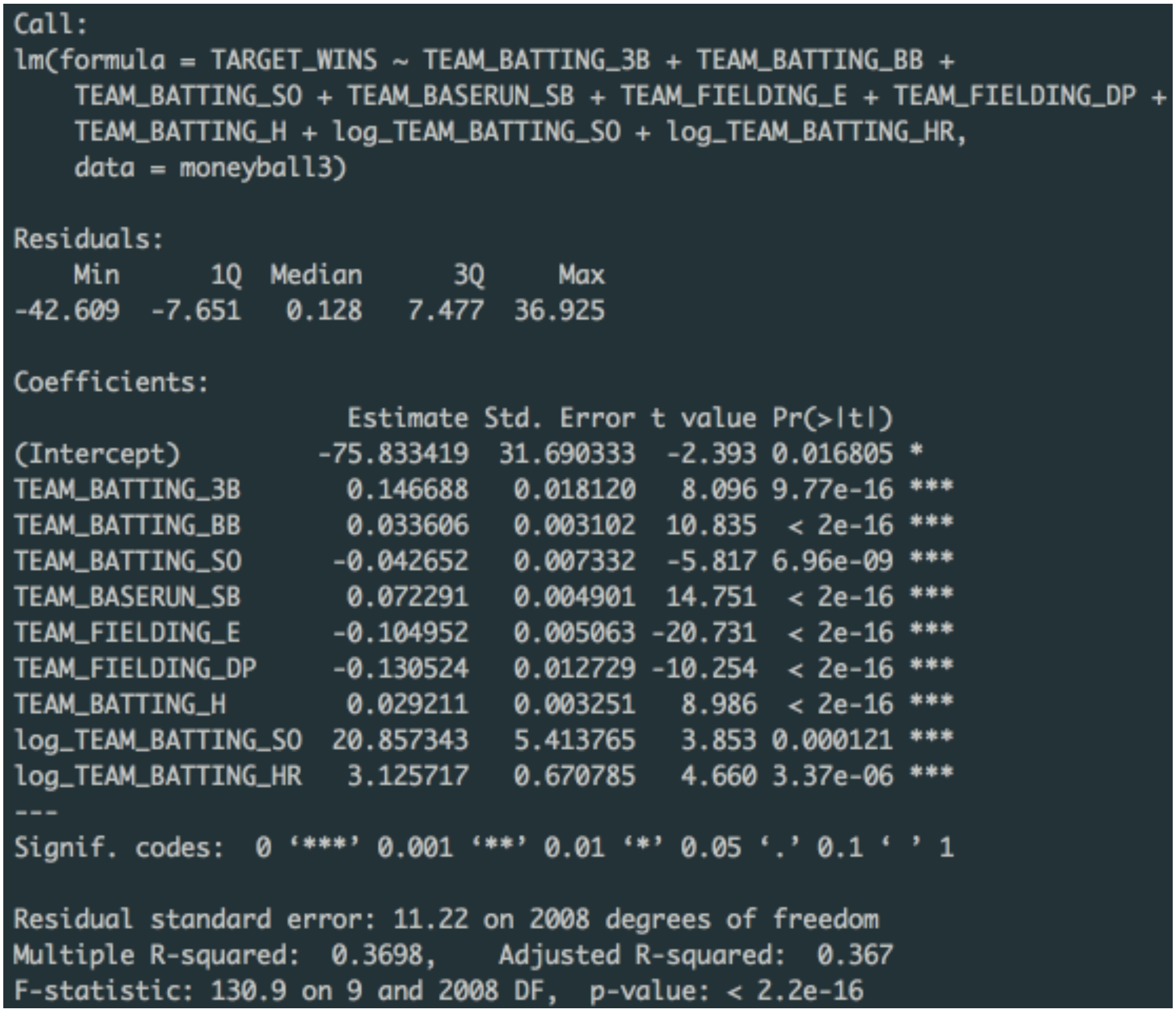
- TEAM_BATTING_3B
- TEAM_BATTING_BB
- TEAM_BATTING_SO
- TEAM_BASERUN_SB
- TEAM_FIELDING_E

- TEAM_FIELDING_DP
- TEAM_BATTING_H
- log_TEAM_BATTING_SO
- log_TEAM_BATTING_HR
- log_TEAM_BASERUN_CS

Based on this assessment using automated variable selection, we then plugged our variables against the response variable into a stepwise model.

Four different models were built in order to obtain the best one. The first model was a stepwise model. Figure 11 outlines the summary output in R for the stepwise model.

Figure 11

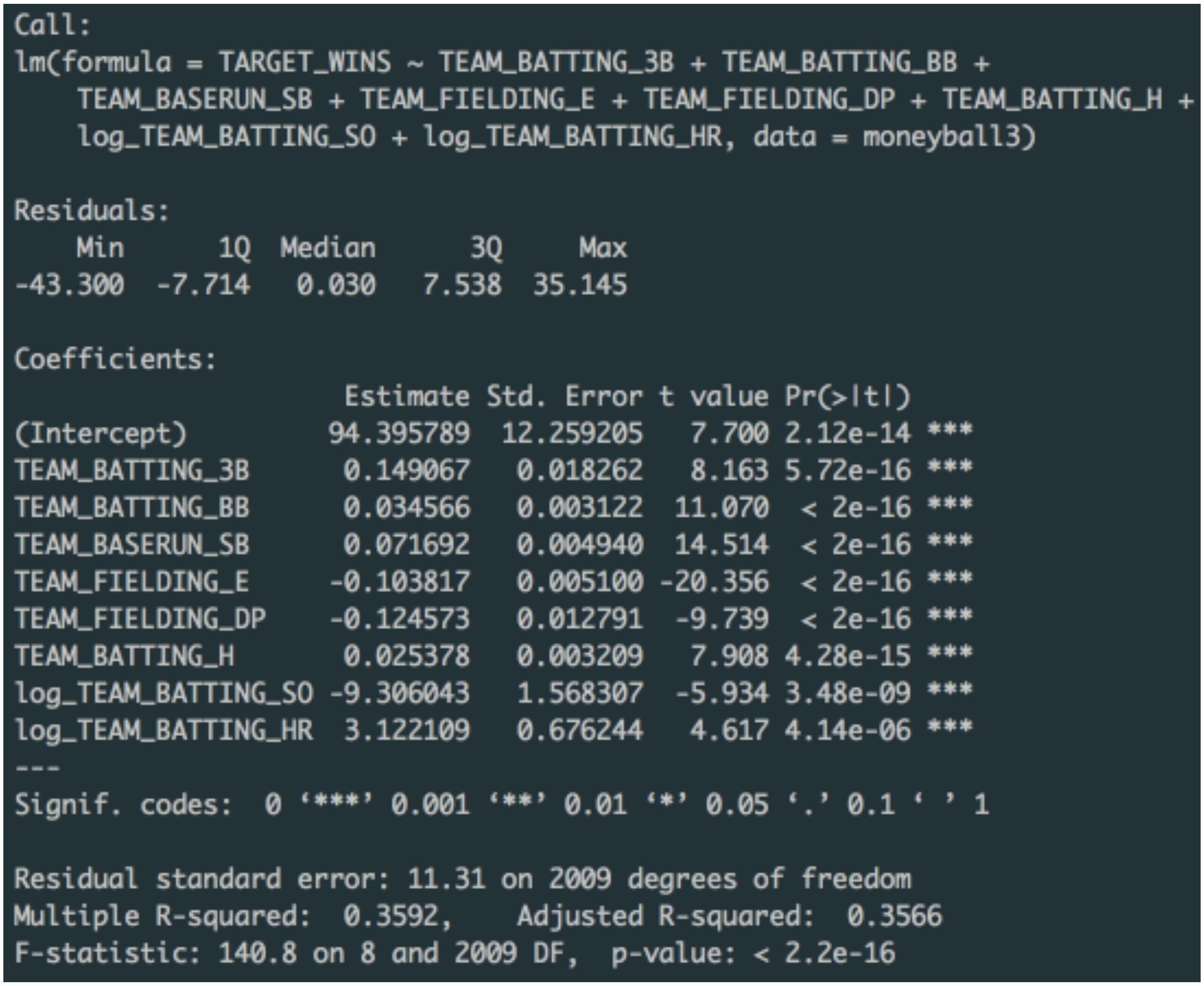


```
Call:
lm(formula = TARGET_WINS ~ TEAM_BATTING_3B + TEAM_BATTING_BB +
    TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_FIELDING_E + TEAM_FIELDING_DP +
    TEAM_BATTING_H + log_TEAM_BATTING_SO + log_TEAM_BATTING_HR,
    data = moneyball3)

Residuals:
    Min      1Q  Median      3Q     Max 
-42.609  -7.651   0.128   7.477  36.925 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -75.833419  31.690333  -2.393 0.016805 *  
TEAM_BATTING_3B       0.146688   0.018120   8.096 9.77e-16 ***
TEAM_BATTING_BB       0.033606   0.003102  10.835  < 2e-16 ***
TEAM_BATTING_SO      -0.042652   0.007332  -5.817 6.96e-09 ***
TEAM_BASERUN_SB       0.072291   0.004901  14.751  < 2e-16 ***
TEAM_FIELDING_E      -0.104952   0.005063 -20.731  < 2e-16 ***
TEAM_FIELDING_DP     -0.130524   0.012729 -10.254  < 2e-16 ***
TEAM_BATTING_H        0.029211   0.003251   8.986  < 2e-16 ***
log_TEAM_BATTING_SO  20.857343   5.413765   3.853 0.000121 ***
log_TEAM_BATTING_HR   3.125717   0.670785   4.660 3.37e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.22 on 2008 degrees of freedom
Multiple R-squared:  0.3698,    Adjusted R-squared:  0.367 
F-statistic: 130.9 on 9 and 2008 DF,  p-value: < 2.2e-16
```

As shown in above, our model had an Adjusted R^2 of 0.367. However, immediately there is an error as our intercept value is a very negative number (-75). The formula is also offset by the log of batting strikeouts and the log of batting home runs. This places heavy emphasis on these two variables.

Our second model utilizes forward selection. Since our first model was not optimal, the decision was to remove the variable that did not score as well as the others from our automated variable selection. This was the log_TEAM_BASERUN_CS variable. There was also a redundancy in variables when evaluated both TEAM_BATTING_SO and log_TEAM_BATTING_SO. For this model, the decision is to remove these values. Figure 12 shows the summary output in R.

Figure 12

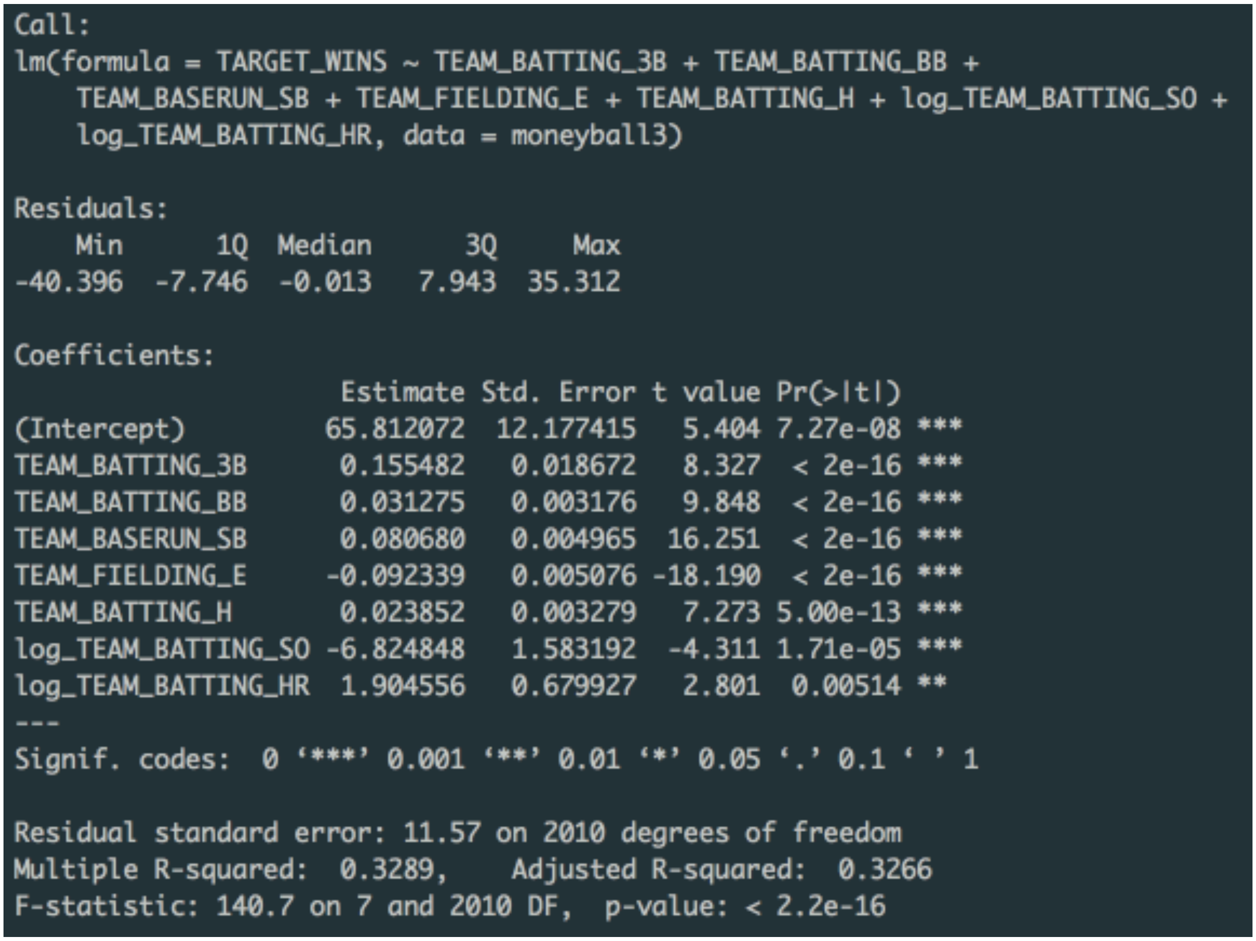


```
Call:
lm(formula = TARGET_WINS ~ TEAM_BATTING_3B + TEAM_BATTING_BB +
    TEAM_BASERUN_SB + TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_BATTING_H +
    log_TEAM_BATTING_SO + log_TEAM_BATTING_HR, data = moneyball3)

Residuals:
    Min      1Q  Median      3Q     Max 
-43.300  -7.714   0.030   7.538  35.145 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         94.395789  12.259205   7.700 2.12e-14 ***
TEAM_BATTING_3B      0.149067   0.018262   8.163 5.72e-16 ***
TEAM_BATTING_BB      0.034566   0.003122  11.070  < 2e-16 ***
TEAM_BASERUN_SB      0.071692   0.004940  14.514  < 2e-16 ***
TEAM_FIELDING_E     -0.103817   0.005100 -20.356  < 2e-16 ***
TEAM_FIELDING_DP    -0.124573   0.012791  -9.739  < 2e-16 ***
TEAM_BATTING_H       0.025378   0.003209   7.908 4.28e-15 ***
log_TEAM_BATTING_SO -9.306043   1.568307  -5.934 3.48e-09 ***
log_TEAM_BATTING_HR  3.122109   0.676244   4.617 4.14e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.31 on 2009 degrees of freedom
Multiple R-squared:  0.3592,	Adjusted R-squared:  0.3566 
F-statistic: 140.8 on 8 and 2009 DF,  p-value: < 2.2e-16
```

The forward selection model scored a bit less in than our stepwise model with an Adjusted R^2 of 0.3566. Although, this model placed a negative value on log_TEAM_FIELDING_DP. Logically, a team that has double plays should have a positive impact on wins.

This leads to our third model; stepwise method. However, this time we only utilized the optimal variables as shown in model 2. Figure 13 below outlines the coefficients of the data using a stepwise method with the most optimal predictor variables.

Figure 13



```
Call:
lm(formula = TARGET_WINS ~ TEAM_BATTING_3B + TEAM_BATTING_BB +
    TEAM_BASERUN_SB + TEAM_FIELDING_E + TEAM_BATTING_H + log_TEAM_BATTING_SO +
    log_TEAM_BATTING_HR, data = moneyball3)

Residuals:
    Min      1Q  Median      3Q     Max
-40.396  -7.746  -0.013   7.943  35.312

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         65.812072  12.177415   5.404 7.27e-08 ***
TEAM_BATTING_3B      0.155482   0.018672   8.327  < 2e-16 ***
TEAM_BATTING_BB      0.031275   0.003176   9.848  < 2e-16 ***
TEAM_BASERUN_SB      0.080680   0.004965  16.251  < 2e-16 ***
TEAM_FIELDING_E     -0.092339   0.005076 -18.190  < 2e-16 ***
TEAM_BATTING_H       0.023852   0.003279   7.273 5.00e-13 ***
log_TEAM_BATTING_SO -6.824848   1.583192  -4.311 1.71e-05 ***
log_TEAM_BATTING_HR  1.904556   0.679927   2.801  0.00514 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.57 on 2010 degrees of freedom
Multiple R-squared:  0.3289,    Adjusted R-squared:  0.3266
F-statistic: 140.7 on 7 and 2010 DF,  p-value: < 2.2e-16
```

As shown above, the coefficients are somewhat different than the forward selection model. The variable coefficients also make logical sense in model 3, it predicts that a given team can win 65 games, with the triples by batters, walks by batters, stolen bases, and walks having a positive impact on wins. Coefficients that are negative include batting strikeouts and errors. The most impactful being strikeouts by batters.

Our final model, was based on my best knowledge of baseball. This model was based, essentially, to level set and compare with the other models. For this model we used the following data:

Variables theoretically that should have a positive impact on wins

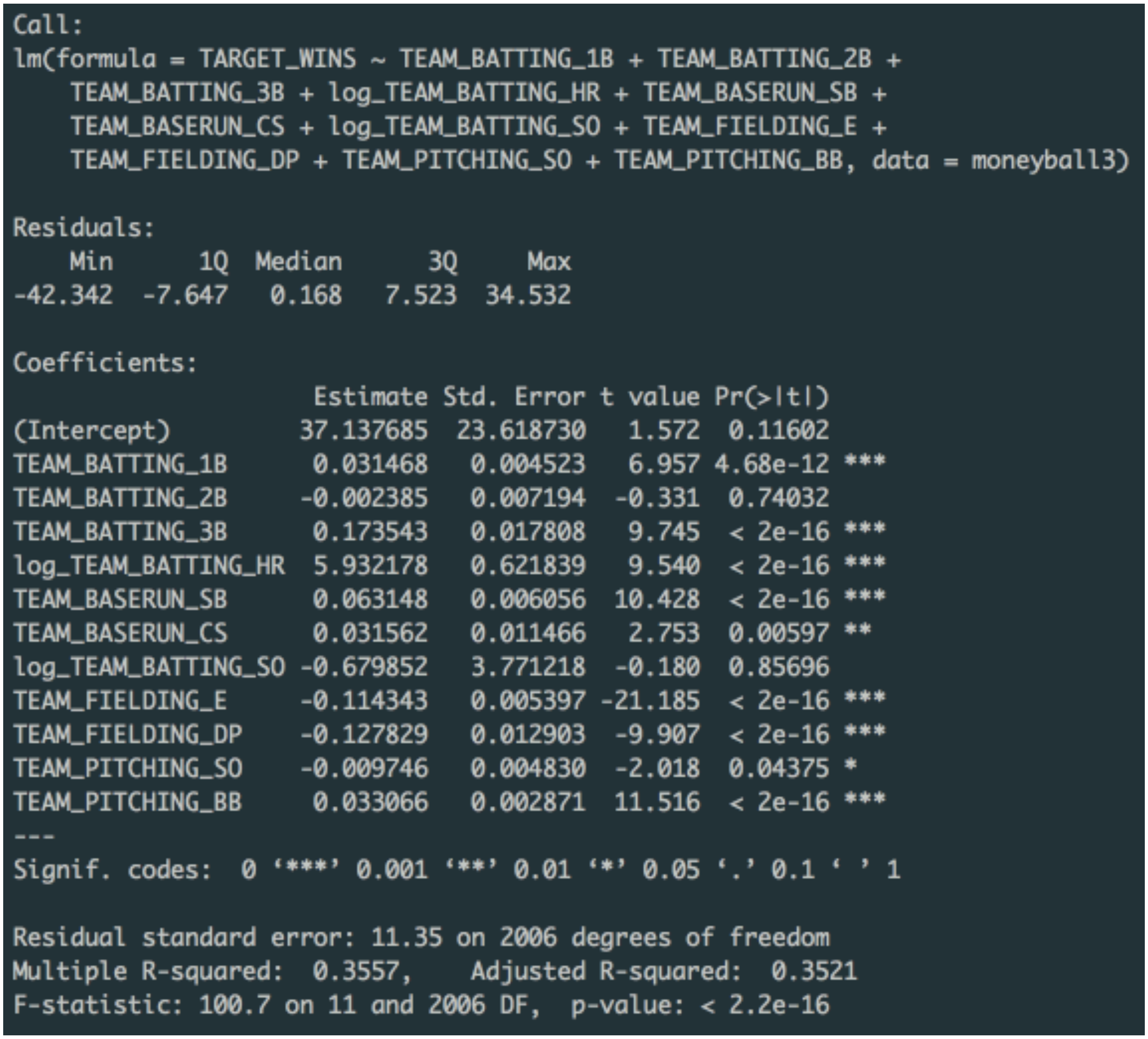
- Singles hit by batters
- Doubles hit by batters
- Triples hit by batters
- Log of batting home runs
- Stolen bases
- Fielding double plays
- Team pitching strike outs

Variables theoretically that should have a negative impact on wins

- Log of team batting strike outs
- Bases caught stealing
- Fielding errors
- Team pitching walks allowed

Based on this knowledge, the items called out to have a theoretical negative impact should have a negative coefficient. Figure 14 below illustrates the coefficients for this model. As shown, fielding errors and team batting strikeouts are negative. These make sense as these theoretically should have a negative impact on wins. However, team batting doubles, team fielding double plays, and team pitching strikeouts all had negative coefficients even though theoretically they should have a positive impact on wins. This model may not be the best to predict our target wins. The Adjusted R^2 for this model also scored lower than other models above at 0.3521.

Figure 14



```
Call:
lm(formula = TARGET_WINS ~ TEAM_BATTING_1B + TEAM_BATTING_2B + 
    TEAM_BATTING_3B + log_TEAM_BATTING_HR + TEAM_BASERUN_SB + 
    TEAM_BASERUN_CS + log_TEAM_BATTING_SO + TEAM_FIELDING_E + 
    TEAM_FIELDING_DP + TEAM_PITCHING_SO + TEAM_PITCHING_BB, data = moneyball3)

Residuals:
    Min      1Q  Median      3Q     Max 
-42.342  -7.647   0.168   7.523  34.532 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         37.137685  23.618730   1.572  0.11602    
TEAM_BATTING_1B      0.031468   0.004523   6.957 4.68e-12 ***
TEAM_BATTING_2B     -0.002385   0.007194  -0.331  0.74032    
TEAM_BATTING_3B      0.173543   0.017808   9.745  < 2e-16 ***
log_TEAM_BATTING_HR  5.932178   0.621839   9.540  < 2e-16 ***
TEAM_BASERUN_SB      0.063148   0.006056  10.428  < 2e-16 ***
TEAM_BASERUN_CS      0.031562   0.011466   2.753  0.00597 ** 
log_TEAM_BATTING_SO -0.679852   3.771218  -0.180  0.85696    
TEAM_FIELDING_E     -0.114343   0.005397 -21.185  < 2e-16 ***
TEAM_FIELDING_DP    -0.127829   0.012903  -9.907  < 2e-16 ***
TEAM_PITCHING_SO    -0.009746   0.004830  -2.018  0.04375 *  
TEAM_PITCHING_BB     0.033066   0.002871  11.516  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.35 on 2006 degrees of freedom
Multiple R-squared:  0.3557,	Adjusted R-squared:  0.3521 
F-statistic: 100.7 on 11 and 2006 DF,  p-value: < 2.2e-16
```

Select Best Model and Stand Alone Scoring:

To asses our model, we will use the AIC, MSE, and Adjusted R^2 statistics. The table below outlines the model metrics.

Model	**AIC**	**MSE**	**Adjusted R^2**
Stepwise (Model 1)	15497.72	125.2077	0.367
Forward (Model 2)	15527.71	127.3346	0.3566
Stepwise Reduced Variables(Model 3)	15618.81	133.3465	0.3266
User Selection (Model 4)	15544.79	128.0352	0.3521

Based on the table, the most optimal should be model 1. Yet, when looking at the coefficients for the variables, the model predicts that a baseball team is looking at losing approximately -75 games before the variables come into play. Obviously this is impossible, the lowest number of games a team can lose is 0.

With this information, the best models are between models 2, 3, and 4. Model 4 scored the highest AIC and the lowers Adjusted R^2. Therefore, we will now look at models 2 and 3. When we created our standalone scoring program on our test data set, we saw that the most optimal model was model 3. This means that based on the predictor variables provided, the average baseball team will win approximately 80 games in a given season. For this to happen, the roster would need to be built with players that specialize in the variables that were presented.

We realize that there are models that have better Adjusted R^2, MSE, and AIC scores. But, in these models, the variable coefficients do not make sense.

After creating our stand alone scoring step in R, the average games a team is predicted to win is 80 games. The scoring values on our test data set did have three errors (index: 300, 436, 1495); these were filled with the mean of the remaining values. Due to the low number of errors, these were deemed to have little effect on the prediction of the model.