DATA 621: Homework 1

Group 5: Christina Valore, Henry Vasquez, Chunhui Zhu, Chunmei Zhu, Yuen Chun Wong

Overview

In this homework assignment, you will explore, analyze and model a data set containing 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 a given year, with all of the statistics adjusted to match the performance of a 162 game season.

Your objective is to build a multiple linear regression model on the training data to predict the number of wins for the team. You can only use the variables given to you (or variables that you derive from the variables provided).

After importing the data sets into GitHub, we further split the training set into two sets to find the RMSE.

Data Exploration

The data set consists of 2,200 records representing the statistics of a team for a season. The seasons range from 1871 to 2006 and include variables for hitting (hits, doubles, triples, home runs, hit-by-pitch, walks and strikeouts), baserunning (stolen bases, caught stealing), fielding (double plays turned, errors made), pitching (walks allowed, hits allowed, home runs allowed, and strikeouts) and the number of wins the team had that season. After running the statistics, we find some of the values to be quite high, for example pitching_h. For a team to allow 16,871 hits for a season of 168 games, means the team allowed approximately 100 hits per game. This value is VERY high, as that would mean the opposing team hit the ball 100 times in 9 innings. There are also some high outliers to take note of in pitching_H and and pitching_SO. This will need to

be addressed in the analysis portion.

```
 TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB
Min.   : 12.00   Min.   :1009   Min.   : 69.0   Min.   :  0.00   Min.   :  0.00   Min.   : 12.0   Min.   :   0.0   Min.   : 14.0
1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00   1st Qu.: 42.00   1st Qu.:451.8   1st Qu.: 550.0   1st Qu.: 66.0
Median : 81.00   Median :1458   Median :238.0   Median : 47.00   Median :102.00   Median :514.0   Median : 752.0   Median :101.0
Mean   : 80.76   Mean   :1469   Mean   :241.3   Mean   : 55.16   Mean   : 99.93   Mean   :504.0   Mean   : 737.6   Mean   :126.2
3rd Qu.: 91.25   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 73.00   3rd Qu.:147.00   3rd Qu.:581.0   3rd Qu.: 930.0   3rd Qu.:158.5
Max.   :135.00   Max.   :2554   Max.   :458.0   Max.   :200.00   Max.   :264.00   Max.   :878.0   Max.   :1335.0   Max.   :697.0
                                                                                                  NA's   :79       NA's   :97
TEAM_BASERUN_CS  TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
Min.   :  7.00   Min.   :29.00    Min.   : 1168   Min.   :  0.0    Min.   : 119.0   Min.   :    0.0   Min.   :  65.0
1st Qu.: 38.00   1st Qu.:52.00    1st Qu.: 1419   1st Qu.: 49.0    1st Qu.: 478.0   1st Qu.:  615.0   1st Qu.: 127.0
Median : 49.00   Median :59.00    Median : 1519   Median :107.0    Median : 537.5   Median :  816.0   Median : 158.5
Mean   : 52.94   Mean   :60.42    Mean   : 1753   Mean   :105.7    Mean   : 554.4   Mean   :  814.4   Mean   : 246.1
3rd Qu.: 62.00   3rd Qu.:69.00    3rd Qu.: 1681   3rd Qu.:151.0    3rd Qu.: 611.0   3rd Qu.:  966.0   3rd Qu.: 249.0
Max.   :201.00   Max.   :95.00    Max.   :16871   Max.   :320.0    Max.   :3645.0   Max.   :12758.0   Max.   :1898.0
NA's   :583      NA's   :1567                                                       NA's   :79
TEAM_FIELDING_DP
Min.   : 64.0
1st Qu.:131.0
Median :148.0
Mean   :146.4
3rd Qu.:163.0
Max.   :228.0
NA's   :217
```

The variables for caught stealing and hit-by-pitch (TEAM_BASERUN_CS and TEAM_BATTING_HBP) were removed because a majority of their values were missing, over 20%. As you can see below the percentage of NA's (p_na) is 34.13% for baserun_CS and batting_HBP for batting_HBP.

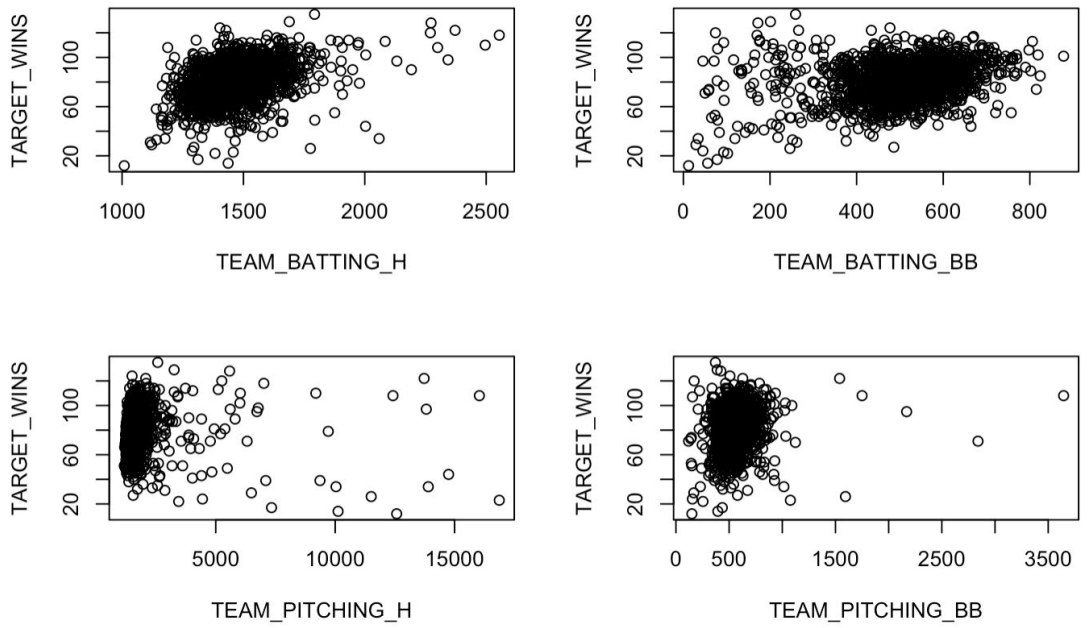
variable <chr>	q_zeros <int>	p_zeros <dbl>	q_na <int>	p_na <dbl>	q_inf <int>	p_inf <dbl>	type <fctr>	unique <int>
TARGET_WINS	0	0.00	0	0.00	0	0	integer	102
TEAM_BATTING_H	0	0.00	0	0.00	0	0	integer	523
TEAM_BATTING_2B	0	0.00	0	0.00	0	0	integer	229
TEAM_BATTING_3B	1	0.06	0	0.00	0	0	integer	138
TEAM_BATTING_HR	11	0.64	0	0.00	0	0	integer	241
TEAM_BATTING_BB	0	0.00	0	0.00	0	0	integer	485
TEAM_BATTING_SO	14	0.82	79	4.63	0	0	integer	745
TEAM_BASERUN_SB	0	0.00	97	5.68	0	0	integer	326
TEAM_BASERUN_CS	0	0.00	583	34.13	0	0	integer	120
TEAM_BATTING_HBP	0	0.00	1567	91.74	0	0	integer	51

We also pay attention to the p_zeros or percentage of zeros, and they all fall below 1%, so those variables are acceptable to keep.

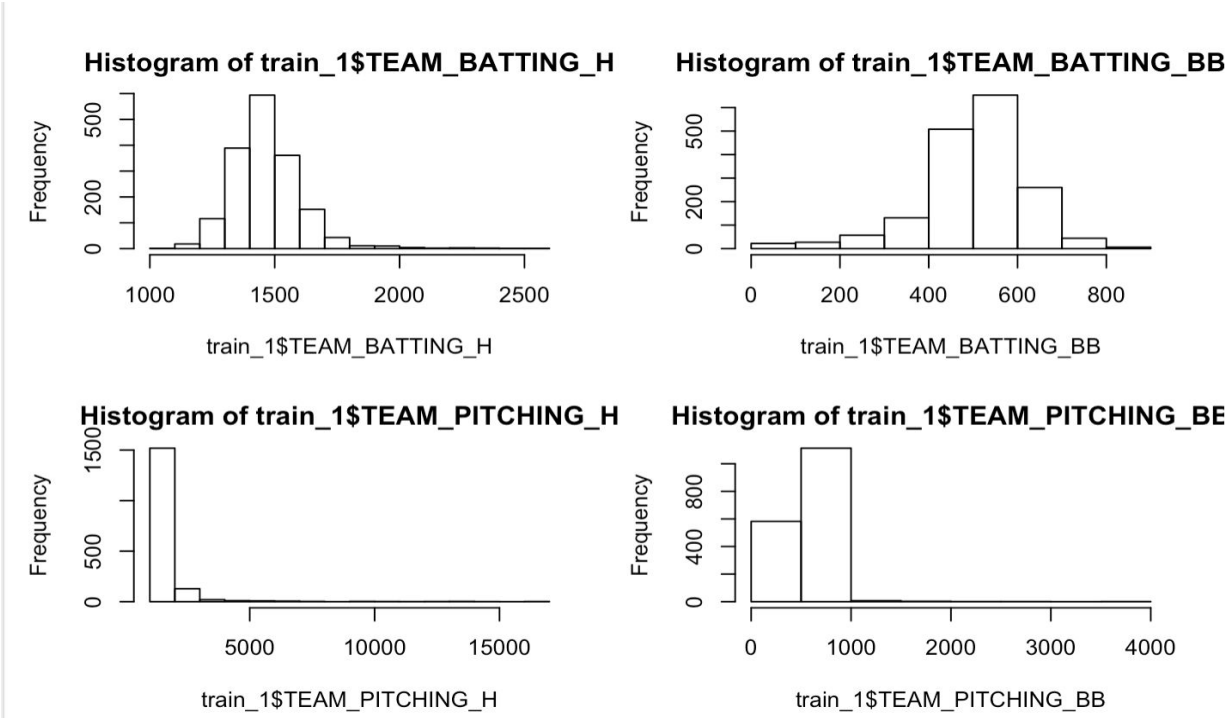
After some research, we decided to investigate the hits, walks and hits allowed and walks allowed as we think these four variables overall will have an impact on the wins. By examining the scatter plots, we see that team batting base hits and team batting walks seem to have a linear relationship to wins as those variables increase it looks as if wins increase as well.

Team pitching hits allowed and team pitching walks allowed, are not as clear and may not have a linear relationship to wins. They both also look to have more extreme outliers as we saw in the

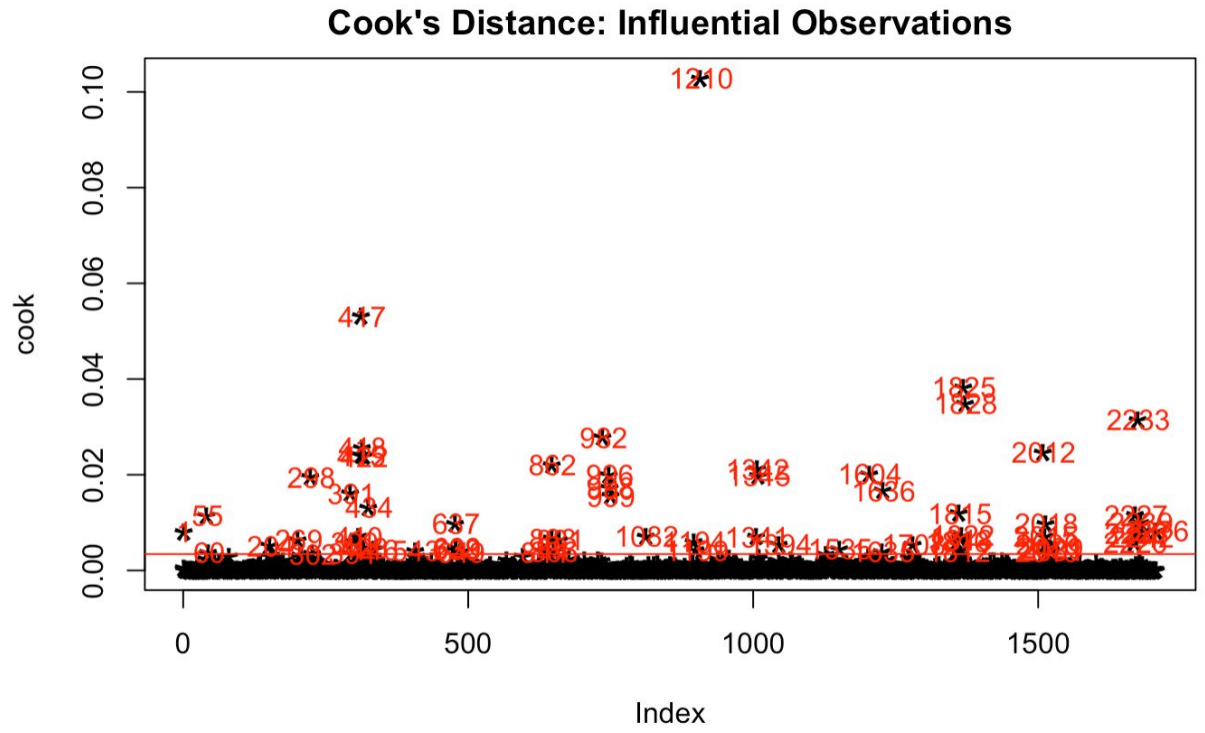
summary.



Taking a look at the histograms next, Batting_H and Batting_BB look to be the most normally distributed, where pitching_H and pitching_BB seem to be more concentrated to the left again indicating there may not be a linear relationship to target wins.



Since this will be a multi-variable regression, we want to consider how all four variables will affect the wins. This is done by using Cook's distance, which calculates the influence each point (row) has on the predicted outcome.



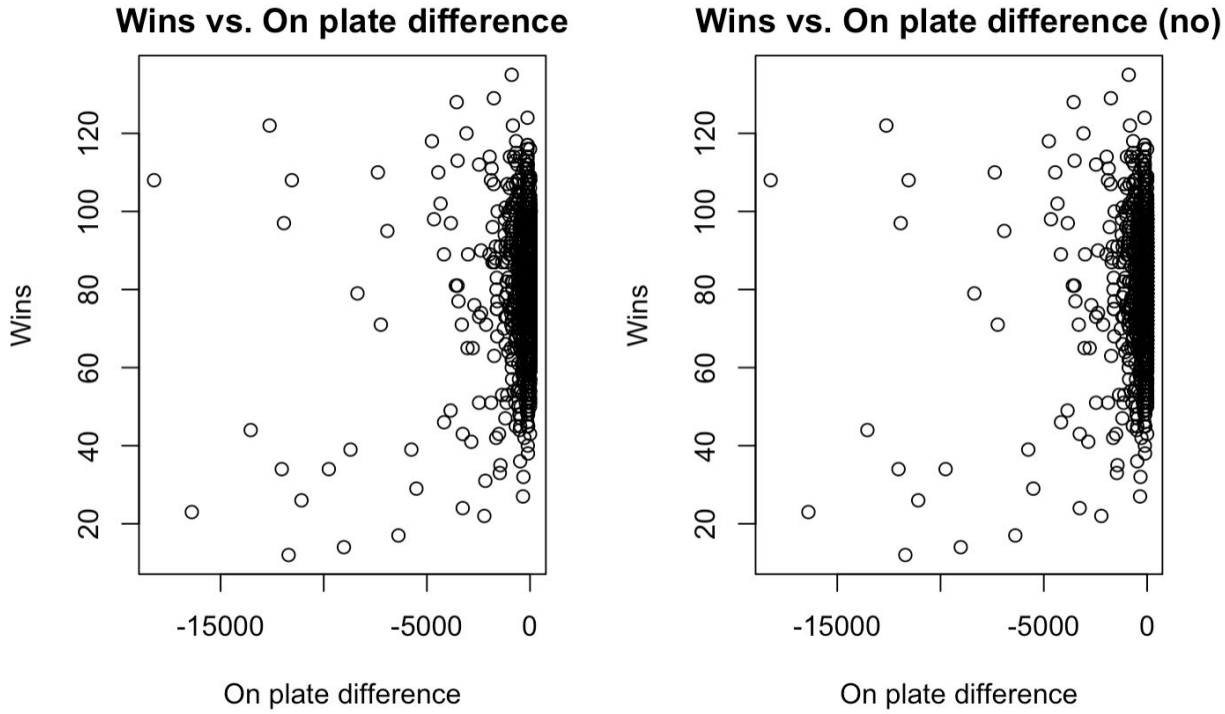
We have 66 influential outliers, with the most influential outlier being at row 1828. In the next section we will discuss how we handle outliers.

Data Preparation

For the missing values, we imputed the median on the training dataset. With regards to the outliers, we decided to use two test sets: one with outliers and one without. Once we tested the variation against out models, we will decide which set to use for the test. We then created three variables:

- team batting on base = team batting base hits + team batting walks
- team pitching allows = team pitching hits allowed + team pitching walks allowed
- BASE_DIFF = team batting on base - team pitching allows

Our thought is that the team that is on base more, will have a higher chance of winning. After creating our new variable base_diff, we plot the value against total wins. The graph to the left have outliers and the one to the right is without. The relationship does not look linear, so perhaps in our models we will need to include more variables.



Build Models

For the project, 4 models were built using the train data set including the outliers, and 4 for the train data set excluding the outliers. The variables chosen were those that did not correlate with one another, for example TEAM_BATTING_H correlates with TEAM_BATTING_2B and TEAM_BATTING_3B so only one of the 3 would have been used.

A statistic that is important in the sport of baseball is OBP, which means on-base-percentage. This is the percentage of times a player reaches base in any given at-bat. However, opportunities or at-bats were not included in the data, so as an alternative, TEAM_BATTING_OB (TEAM_BATTING_H + TEAM_BATTING_BB) was created which attempts to calculate the number of times the team had runners reach base in a given season. The motive behind including this variable is that a team with more opportunities to score tends to score more, and therefore win more.

Defensive aspects were also included in all of the models, such as TEAM_FIELDING_E and TEAM_FIELDING_DP. Teams that make errors allow the opposing team more opportunities to score and result in less wins in a given season. Teams that also turn a lot of double plays suggest good defense, however it can also suggest that a team gives up a lot of singles, base-on-balls, or makes a lot of errors. Regardless, double-plays play a role in defense and can lead to less runs allowed.

Since pitching is another big part of the game, pitching variables were included in the models. Similar to the variable TEAM_BATTING_OB that was created, another variable called TEAM_PITCHING_A was introduced to estimated the total base runners that pitchers allowed. This variable is the sum of TEAM_PITCHING_H and TEAM_PITCHING_BB. The idea is that teams that allow opponents to reach base more give up more runs, which results in more loses.

Model 1

In model 1, 2 of the estimators used showed no significance in their p-values, as well as TEAM_FIELDING_DP having a negative value. Double plays are a positive for a team, which should result in a positive coefficient. However, as mentioned previously it could suggest that the team gave up a lot of baserunners which would explain its negative value.

Model 2

Model 2 was based off model 1 but removing the estimator TEAM_PITCHING_A. The result was all of the estimators were significant, but TEAM_FIELDING_DP with a contradicting coefficient. Aside from the contradicting coefficient, the remaining estimators had coefficients that aligned with their impact on team performance.

Model 3

Model 3 removes the contradicting coefficient from model 2. The result is a model with all estimators having significant p-values, and coefficients with correct positive and negative values. More OB suggesting more offense, which results in more wins. And more fielding errors, resulting in more opportunities for opponents, and results in more losses.

Model 4

Model 4, which does not have the highest r-squared was deemed the best model of the 4. It included the coefficients in model 3 with the addition of stolen bases. All of the estimators have significant p-values,

and the r-squared is not much lower than models 1 and 2.The coefficients for each estimator also corresponds with the proper outcome for the team.

Conclusions on the models

Although some of the models had estimator with contradicting coefficients, they were kept for further assessment.

Select Models

Multicollinearity was checked for all of the models, and it resulted that none of the models had any serious multicollinearity issues. After testing each model for the assumptions of linear regression all models passed. There were no models with heteroskedasticity issues or non-linearity problems.

The model selected for the project was model 4. It had an r-squared that was similar to models 1 and 2, which had the highest r-squared but also had estimators with coefficients that aligned with the impact on the team. TEAM_BATTING_OB and TEAM_BASERUN_SB both positive attributes for a team have positive coefficients, and TEAM_FIELDING_E a negative attribute had a negative coefficient.

In models 1 and 2, the estimator TEAM_FIELDING_DP, a positive attribute for the team resulted in a negative coefficient. It was previously explained that this could be the cause of pitchers allowing more walks or singles, or also teams making for errors. However, since the data is limited, these 2 models were removed from consideration.

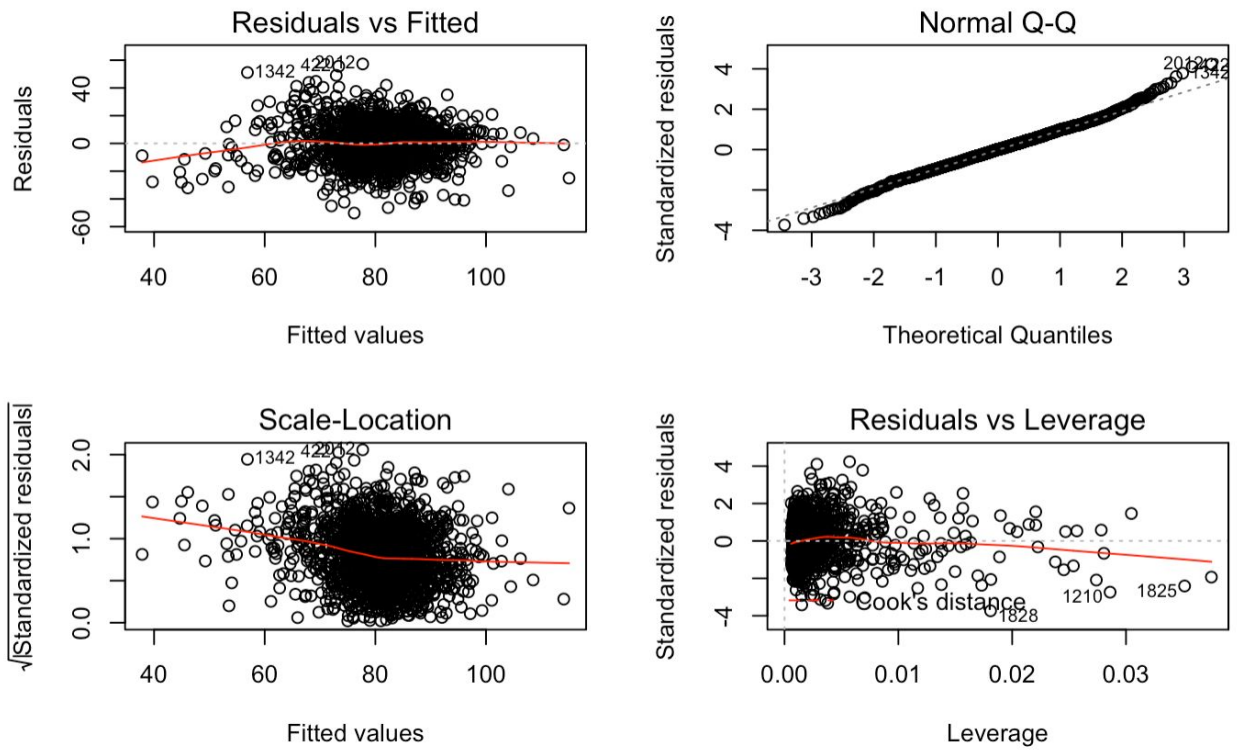
There were no issues with model 3, however since model 4 was created by adding a variable to model 3 and resulted in a slightly better performance, model 4 was chosen:

```
Call:
lm(formula = TARGET_WINS ~ TEAM_BATTING_OB + TEAM_FIELDING_E +
    TEAM_BASERUN_SB, data = train_no)

Residuals:
    Min      1Q  Median      3Q     Max
-50.558  -8.795   0.099   8.562  57.158

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.515209   3.803670   2.239   0.0253 *
TEAM_BATTING_OB 0.036288   0.001888  19.224  < 2e-16 ***
TEAM_FIELDING_E -0.009534   0.001635  -5.832 6.55e-09 ***
TEAM_BASERUN_SB 0.024093   0.004095   5.884 4.83e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.57 on 1659 degrees of freedom
Multiple R-squared:  0.2381,    Adjusted R-squared:  0.2367
F-statistic: 172.8 on 3 and 1659 DF,  p-value: < 2.2e-16
```



Please see appendix for additional tests such as multicollinearity, RMSE, mean squared error r-squared, F-statistic, and residual plots of the models.

Finally, using the evaluation dataset, we were able to make predictions for future games. The .CSV can be seen here:
https://raw.githubusercontent.com/ChristinaValore/Business-Analytics-and-Data-Mining-621/master/Homework1/Predictions_Baseball.csv

Final Thoughts

The data set included teams from 1871-2006, which brings up multiple issues. For starters, some wins in a season are estimated because early seasons did not actually have 162 games. In fact, the 1871 season had only 29 games.

Teams that play the full 162 game season tend to normalize as the get later into the season. This is because they go through hot and cold streaks, where they win 10/10 games or lost 9/10 games. In some seasons, lock outs did not result in a full baseball season. For example, in 1972 the lockout cancelled 86 games.

Because of these inconsistencies in the season lengths, some of the estimated variables are extremely high or extremely low. An example can be seen in the max TEAM_PITCHING_H in

the data which is 30,132. This suggests that a team gave up an average of 186 hits per game, which is not possible. Similarly the max for TEAM_PITCHING_SO is 19,278. This is an average of 119 strikeouts per game, when there's only 27 outs in a game.

The models could perform better with data that stays consistent to an era, or that stays true to teams that actually played 162 games in a season.

Appendix

https://github.com/ChristinaValore/Business-Analytics-and-Data-Mining-621/blob/master/data621_assignment1.Rmd