IS621_hw2

Charley Ferrari

February 16, 2016

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Data Exploration

The moneyball dataset includes 2276 observations of 17 variables. I will be looking at WINS as my dependent variable. Discounting the ID variable, I will be using the other 14 variables as possible independent variables.

Below is a table describing the variables and expected effect on wins:

VARIABLE_NAME	DEFINITION	THEORETICAL_EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	Target
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	Positive Impact on Wins
TEAM_BATTING_2B	Doubles by batters (2B)	Positive Impact on Wins
TEAM_BATTING_3B	Triples by batters (3B)	Positive Impact on Wins
TEAM_BATTING_HR	Homeruns by batters (4B)	Positive Impact on Wins
TEAM_BATTING_BB	Walks by batters	Positive Impact on Wins
TEAM_BATTING_HBP	Batters hit by pitch (get a free base)	Positive Impact on Wins
TEAM_BATTING_SO	Strikeouts by batters	Negative Impact on Wins
TEAM_BASERUN_SB	Stolen bases	Positive Impact on Wins
TEAM_BASERUN_CS	Caught stealing	Negative Impact on Wins
TEAM_FIELDING_E	Errors	Negative Impact on Wins
TEAM_FIELDING_DP	Double Plays	Positive Impact on Wins
TEAM_PITCHING_BB	Walks allowed	Negative Impact on Wins
TEAM_PITCHING_H	Hits allowed	Negative Impact on Wins
TEAM_PITCHING_HR	Homeruns allowed	Negative Impact on Wins
TEAM_PITCHING_SO	Strikeouts by pitchers	Positive Impact on Wins

With 2276 rows of 17 variables, one would expect to have 38692 records. However, some values are missing, leaving 35214 actual records.

Below is a summary of the missing variables, starting with the Dependent variable TARGET_WINS:

```
## TEAM_BATTING_SO TEAM_BASERUN_SB TEAM_BASERUN_CS
## "NA's   :102  " "NA's   :131  " "NA's   :772  "

## TEAM_BATTING_HBP TEAM_PITCHING_SO TEAM_FIELDING_DP
## "NA's   :2085  "  "NA's   :102  "  "NA's   :286  "
```

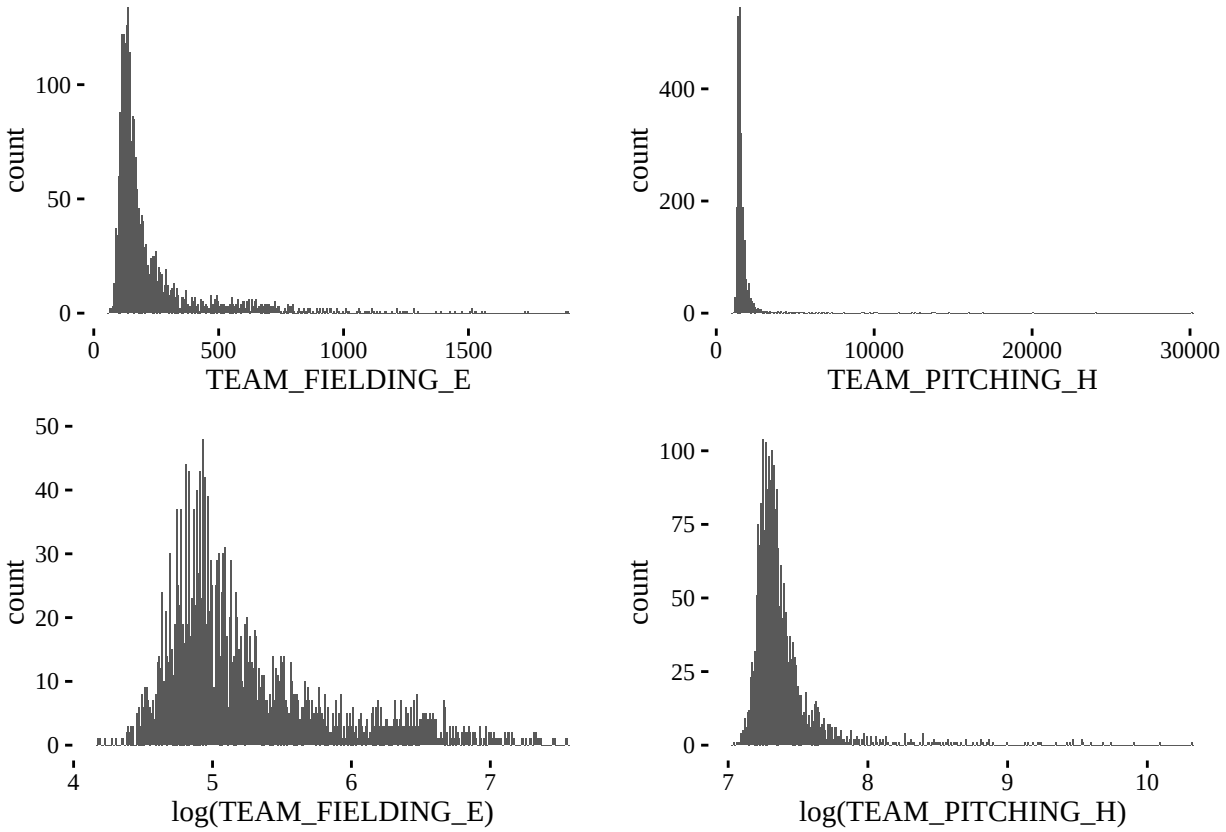
Out of 2276 total observations, there are 191 with at least one variable missing, or about 8.4% of the rows. I'll try versions of models with both the missing values excluded, and with means of the variables replacing missing values for comparison purposes. Because we haven't picked our model yet, I don't think it would make sense to run a regression to replace these missing values. If we had a model in mind already, it would make more sense.

To summarize the data I'll look at boxplots, and check out scatter plots of each variable versus the target variable. The (large number of) plots are presented in the Appendix, with the commentary below.

Several variables appeared to have outliers. Looking at the original scatterplots, there were a few variables that seemed to have anomolous outliers, and a few more predictable outliers that suggested a variable transformation was needed.

Data Preparation

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. For this variable, I'll also remove outliers.

In addition, Walks and strikeouts by pitchers seemed to have outliers. I'll remove these by removing all values outside of 3 standard deviations.

Based on the scatterplots of variables vs WINS, the residuals of Homeruns by batters, walks by batters, and homeruns allowed by pitchers seem potentially heteroskedastic. Right now I'll perform no transformation to correct for this, but will keep this in mind for variable selection.

Now, lets take a look at a correlation matrix of our variables:

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
3	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0
4	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	1	1	0	1	0	0	0	1	1	0	1	1	0
6	0	0	1	1	0	1	0	0	0	0	1	0	1	1	0
7	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
8	0	0	1	1	0	1	0	0	0	1	1	0	1	1	0
9	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
10	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
12	1	0	1	0	0	1	0	0	0	1	0	0	0	1	0
13	0	0	1	1	0	1	0	0	0	0	1	0	1	1	0
14	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
15	0	0	1	1	0	1	0	0	0	0	1	0	1	0	0
16	0	0	1	1	0	1	1	0	0	1	1	0	0	1	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

To display the entire matrix, I just used 1 if the correlation is greater than 0.5 or less than -0.5, and 0 otherwise. Some correlations make sense, the correlations between batters who hit versus the different types of bases for example. For the moment, I will use this as a rule of thumb of which variables to pick, and won't try to capture possible interaction effects.

I'll also check to see whether the p-values are above 0.05:

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
3	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
6	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
8	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0
11	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
12	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
14	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0
15	0	0	0	0	1	0	0	0	1	0	0	1	0	0	1
16	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0

Looking at single variable models, lets look at our adjusted r-squares with missing rows removed:

variable	adj_r_squared
TEAM_BATTING_H	0.1487504
TEAM_BATTING_2B	0.0581076

variable	adj_r_squared
TEAM_PITCHING_BB	0.0567488
TEAM_BATTING_BB	0.0542675
TEAM_PITCHING_H	0.0362713
TEAM_PITCHING_HR	0.0361937
TEAM_BATTING_HR	0.0329524
TEAM_BATTING_3B	0.0165104
TEAM_FIELDING_E	0.0127851
TEAM_BASERUN_SB	0.0120583
TEAM_PITCHING_SO	0.0070552
TEAM_BATTING_SO	0.0039982
TEAM_FIELDING_DP	0.0005848
TEAM_BATTING_HBP	0.0001405
TEAM_BASERUN_CS	-0.0005540

And with missing values replaced with means:

variable	adj_r_squared
TEAM_BATTING_H	0.1487504
TEAM_BATTING_2B	0.0581076
TEAM_PITCHING_BB	0.0567488
TEAM_BATTING_BB	0.0542675
TEAM_PITCHING_H	0.0362713
TEAM_PITCHING_HR	0.0361937
TEAM_BATTING_HR	0.0329524
TEAM_BATTING_3B	0.0165104
TEAM_FIELDING_E	0.0127851
TEAM_BASERUN_SB	0.0107856
TEAM_PITCHING_SO	0.0070552
TEAM_BATTING_SO	0.0039982
TEAM_FIELDING_DP	0.0003799
TEAM_BATTING_HBP	-0.0001302
TEAM_BASERUN_CS	-0.0004113

This table will give us an idea of which models to test in our variable selection phase.

Build Models

For this assigmnent I'll use informal forward variable selection. I'll use this method to step through several models, but will use other criteria at times to inform my variable choices.

There are a few common sense decisions I will make in terms of my selections. In particular, I will avoid having variables that are conceptually similar, which might suggest collinearity. For example, I will only use base hits by batters rather than individually pick singles, doubles, triples, or home runs.

The highest variable in both cases (missing values removed and means added for missing values) is total base hits for batters, so this will be my first variable. Now, lets look at adjusted R-squares of a second variable. Here is the data with the missing variables removed:

variable	adj_r_squared
TEAM_BATTING_HBP	0.2202497

variable	adj_r_squared
TEAM_BATTING_BB	0.1868865
TEAM_FIELDING_E	0.1806755
TEAM_BATTING_HR	0.1787896
TEAM_PITCHING_HR	0.1782509
TEAM_PITCHING_BB	0.1764152
TEAM_BATTING_SO	0.1585107
TEAM_PITCHING_SO	0.1535839
TEAM_PITCHING_H	0.1515854
TEAM_BATTING_3B	0.1499264
TEAM_BATTING_2B	0.1486646
TEAM_FIELDING_DP	0.1384582
TEAM_BASERUN_SB	0.1351467
TEAM_BASERUN_CS	0.1266394

And with missing values replaced with means:

variable	adj_r_squared
TEAM_BATTING_BB	0.1868865
TEAM_FIELDING_E	0.1806755
TEAM_BATTING_HR	0.1787896
TEAM_PITCHING_HR	0.1782509
TEAM_PITCHING_BB	0.1764152
TEAM_BATTING_SO	0.1585107
TEAM_FIELDING_DP	0.1542943
TEAM_PITCHING_SO	0.1535839
TEAM_PITCHING_H	0.1515854
TEAM_BASERUN_SB	0.1505796
TEAM_BATTING_3B	0.1499264
TEAM_BATTING_HBP	0.1487763
TEAM_BATTING_2B	0.1486646
TEAM_BASERUN_CS	0.1484944

Batters hit by pitch seems to drastically change depending on our treatment of missing variables, so I'll remove it from consideration for now, meaning walks by batters is getting chosen as my next variable. Both of these variabes are significant, so I'll continue my forward selection.

Once again, with missing values removed:

variable	adj_r_squared
TEAM_BATTING_HBP	0.3622632
TEAM_FIELDING_DP	0.2227758
TEAM_BASERUN_CS	0.2136569
TEAM_BASERUN_SB	0.2029552
TEAM_PITCHING_HR	0.1965258
TEAM_BATTING_HR	0.1961016
TEAM_FIELDING_E	0.1960549
TEAM_PITCHING_SO	0.1902787
TEAM_BATTING_SO	0.1889304
TEAM_PITCHING_H	0.1884580
TEAM_BATTING_3B	0.1874497

variable	adj_r_squared
TEAM_BATTING_2B	0.1871499
TEAM_PITCHING_BB	0.1864975

And missing values replaced by means:

variable	adj_r_squared
TEAM_FIELDING_DP	0.2056875
TEAM_PITCHING_HR	0.1965258
TEAM_BATTING_HR	0.1961016
TEAM_FIELDING_E	0.1960549
TEAM_BASERUN_SB	0.1911546
TEAM_PITCHING_SO	0.1902787
TEAM_BATTING_SO	0.1889304
TEAM_PITCHING_H	0.1884580
TEAM_BATTING_3B	0.1874497
TEAM_BATTING_2B	0.1871499
TEAM_BATTING_HBP	0.1868448
TEAM_BASERUN_CS	0.1867207
TEAM_PITCHING_BB	0.1864975

Once again ignoring batters hit by pitches, Double Plays are now giving us the largest adjusted r-square. All variables are significant, so we'll continue with our additions.

Missing values removed:

variable	adj_r_squared
TEAM_BATTING_HBP	0.3878109
TEAM_FIELDING_E	0.2477493
TEAM_BASERUN_SB	0.2374316
TEAM_BATTING_HR	0.2362692
TEAM_PITCHING_HR	0.2358117
TEAM_BASERUN_CS	0.2334635
TEAM_BATTING_2B	0.2274009
TEAM_BATTING_3B	0.2236856
TEAM_PITCHING_H	0.2231049
TEAM_PITCHING_SO	0.2228667
TEAM_BATTING_SO	0.2227230
TEAM_PITCHING_BB	0.2224159

And missing values replaced by means:

variable	adj_r_squared
TEAM_PITCHING_HR	0.2222335
TEAM_BATTING_HR	0.2216677
TEAM_FIELDING_E	0.2208517
TEAM_PITCHING_SO	0.2103103
TEAM_BATTING_SO	0.2091451
TEAM_BASERUN_SB	0.2066440

variable	adj_r_squared
TEAM_PITCHING_H	0.2061951
TEAM_BATTING_2B	0.2056125
TEAM_BATTING_HBP	0.2055569
TEAM_PITCHING_BB	0.2054503
TEAM_BASERUN_CS	0.2053240
TEAM_BATTING_3B	0.2053213

Now, it looks like our variable orders are different depending on how we treat outliers. Before it seemed like a clear choice to not consider batters hit by pitch since it was the sole variable that was changing position, but now that multiple variables are switching order I'll have to make a more nuanced choice. Errors appears to be the most stable, remaining just below hit by pitch when missing values are removed, and just being beaten out by home run hits when missing values are replaced by means.

Home runs are also not a good choice since we already have hits of any kind in our model, and home runs would be expected conceptually to have collinearity issues with total base hits. This effect isn't showing up in the collinearity matrix however, suggesting we might want to use more advanced methods to deal with different types of base hits.

So, our choice will be errors for this model. Lets see what the next variable will be.

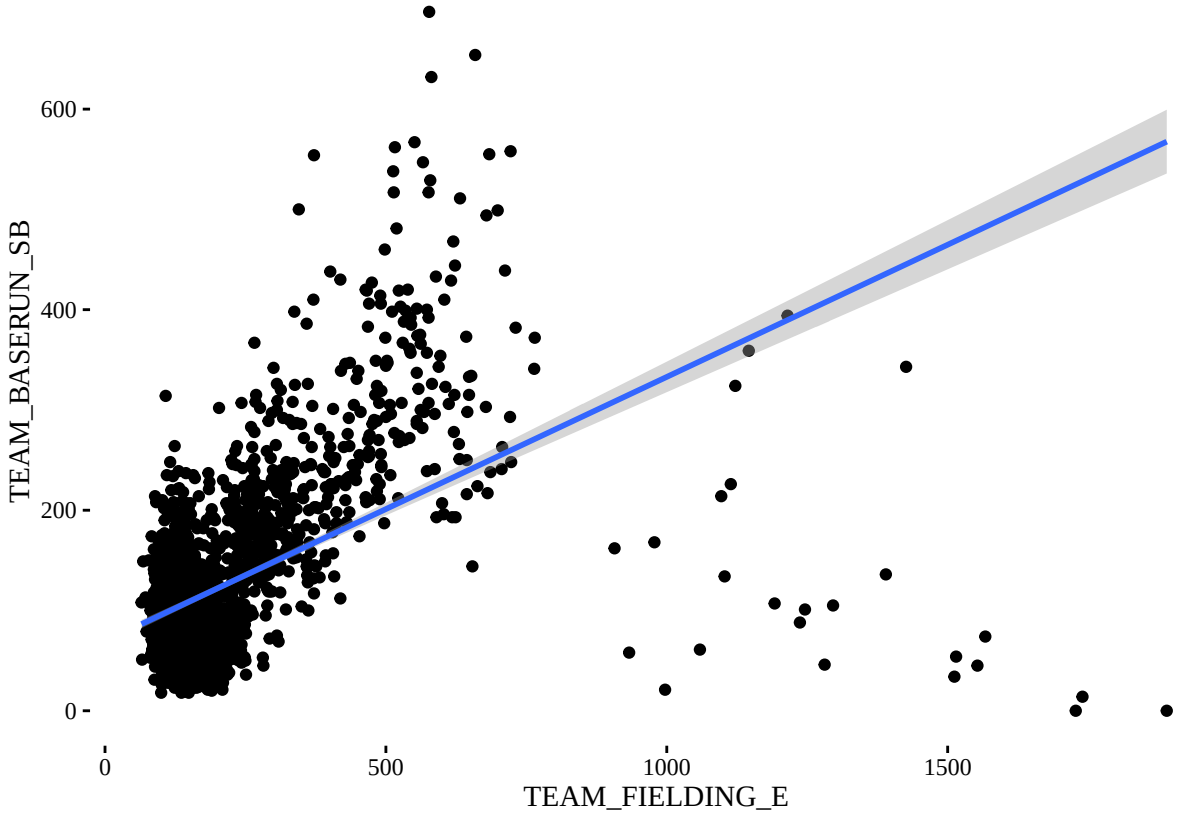
Missing values removed:

variable	adj_r_squared
TEAM_BATTING_HBP	0.4518945
TEAM_BASERUN_CS	0.3098735
TEAM_BASERUN_SB	0.2868785
TEAM_BATTING_3B	0.2765557
TEAM_BATTING_2B	0.2683476
TEAM_BATTING_SO	0.2596148
TEAM_PITCHING_H	0.2592533
TEAM_PITCHING_SO	0.2515637
TEAM_PITCHING_BB	0.2514711
TEAM_PITCHING_HR	0.2479644
TEAM_BATTING_HR	0.2477994

And missing values replaced by means:

variable	adj_r_squared
TEAM_BATTING_3B	0.2366889
TEAM_PITCHING_H	0.2363925
TEAM_BASERUN_SB	0.2360778
TEAM_BATTING_2B	0.2282623
TEAM_PITCHING_BB	0.2251237
TEAM_PITCHING_HR	0.2250926
TEAM_BATTING_HR	0.2241607
TEAM_PITCHING_SO	0.2213444
TEAM_BATTING_HBP	0.2207649
TEAM_BATTING_SO	0.2206230
TEAM_BASERUN_CS	0.2205073

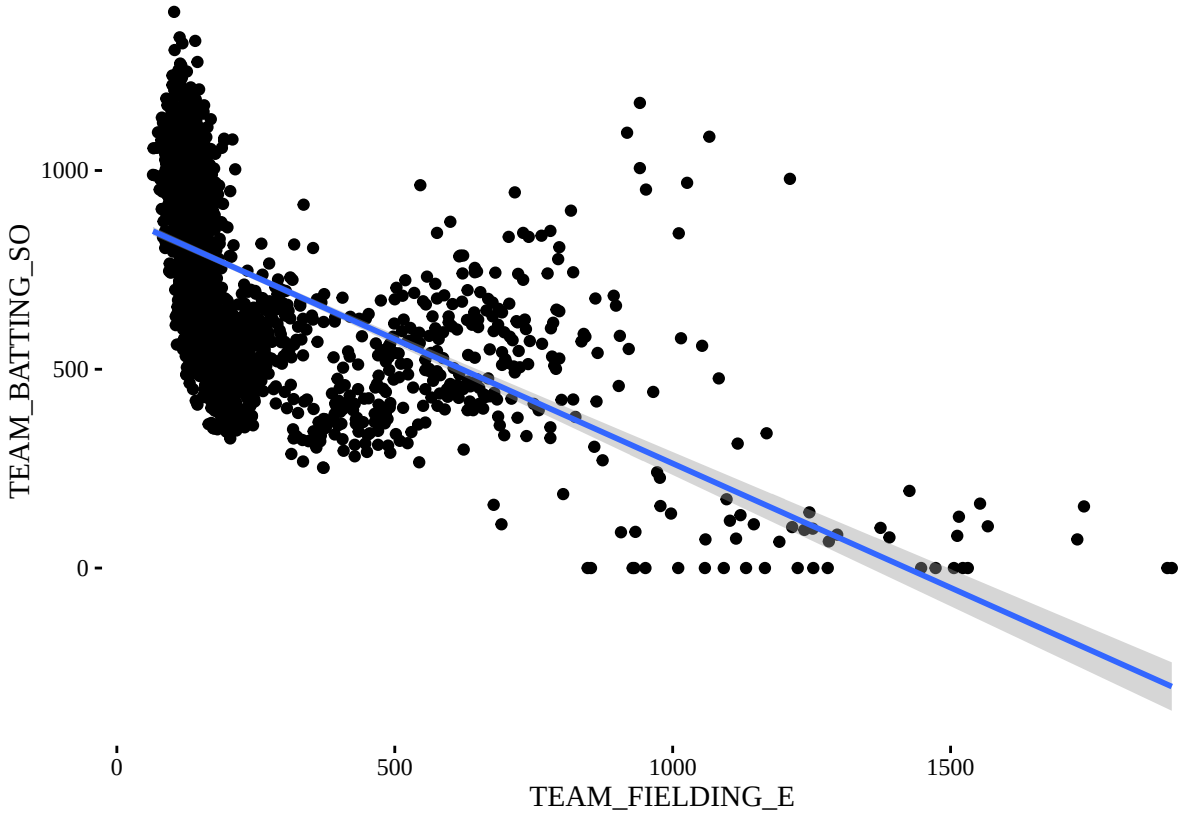
Caught stealing is performing well, but looking back at my missing variables summary, I see that it has 772 records missing. I don't think it would be trustworthy to use this as a variable. Stolen bases is pretty stable near the top, but has a high correlation with errors, which are already in the model. Checking out a scatter plot shows that this relationship is problematic:



There appear to be a few bad leverage points throwing off the overall relationship, meaning there could be a stronger relationship than the correlation matrix is giving us. For this reason, it's probably not a good idea to include stolen bases in my model.

I'm beginning to get diverging results when comparing the method of handling outliers, so for the time being I'm going to stick with picking variables from excluding missing variables, and then taking the number of missing observations of each variables into consideration.

Sticking to this rule, and ignoring variables that for conceptual reasons are correlated (doubles and triples), the best variable to choose is strikeouts by batters. Unfortunately, this variable too has a problematic relationship with errors:



Once again, it looks like we might have groupings of points in this scatterplot. In this case, it suggests we may have multiple relationships, a clear relationship for low error games, and other effects happening when errors are high.

Conceptually, both stolen bases and strikeouts can be seen as types of errors in baseball. It might be more parsimonious to include just errors, in a similar way to just including the number of base hits rather than also including doubles, triples, and homeruns.

I ended up with high correlations for the rest of the variables (mostly with errors.) I tested a few models adding these variables and, while I ended up with p-values less than 0.05, they were in the order of magnitude of 0.01, while the previous p-values were orders of magnitude smaller. This suggests that we were appraoching the end of this algorithm anyway.

Select Model

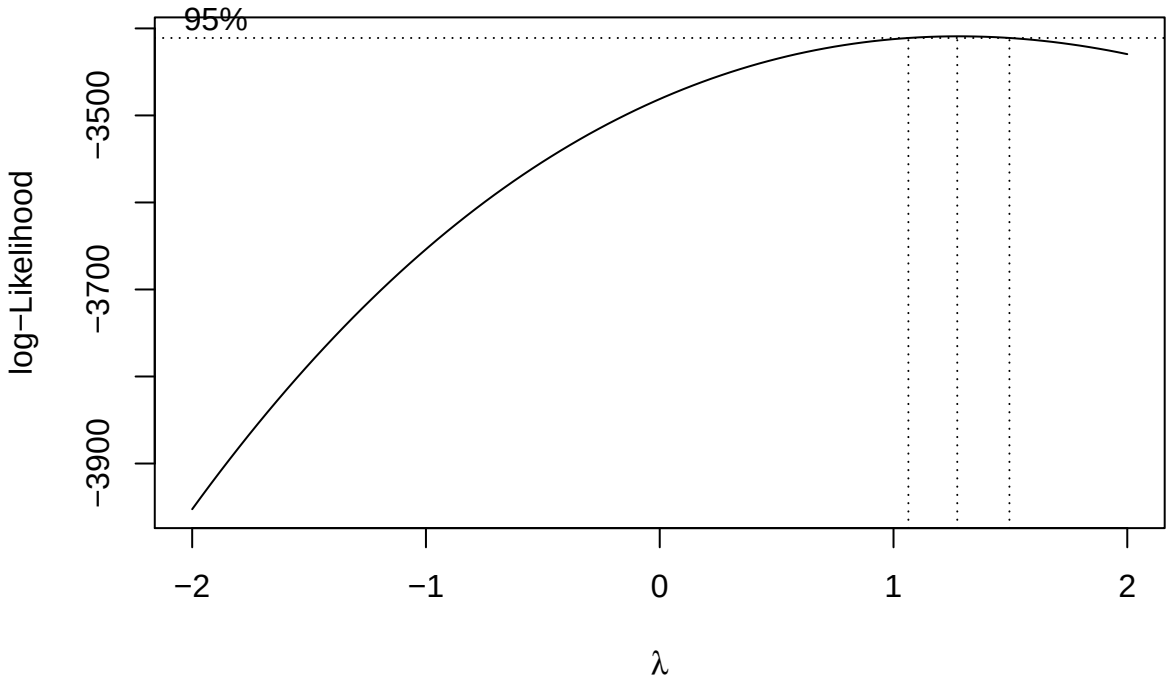
This means that our final model is:

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_BB +
##     TEAM_FIELDING_DP + TEAM_FIELDING_E, data = newmoneyball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.900  -7.928  -0.003   7.925  47.340
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      48.712645   6.398591   7.613 4.23e-14 ***
## TEAM_BATTING_H    0.048418   0.002540  19.061  < 2e-16 ***
## TEAM_BATTING_BB   0.032989   0.003318   9.942  < 2e-16 ***
## TEAM_FIELDING_DP -0.139435   0.012875 -10.830  < 2e-16 ***
## TEAM_FIELDING_E  -6.886843   0.869855  -7.917 4.14e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.57 on 1857 degrees of freedom
##   (232 observations deleted due to missingness)
## Multiple R-squared:  0.2494, Adjusted R-squared:  0.2477
## F-statistic: 154.2 on 4 and 1857 DF,  p-value: < 2.2e-16
```

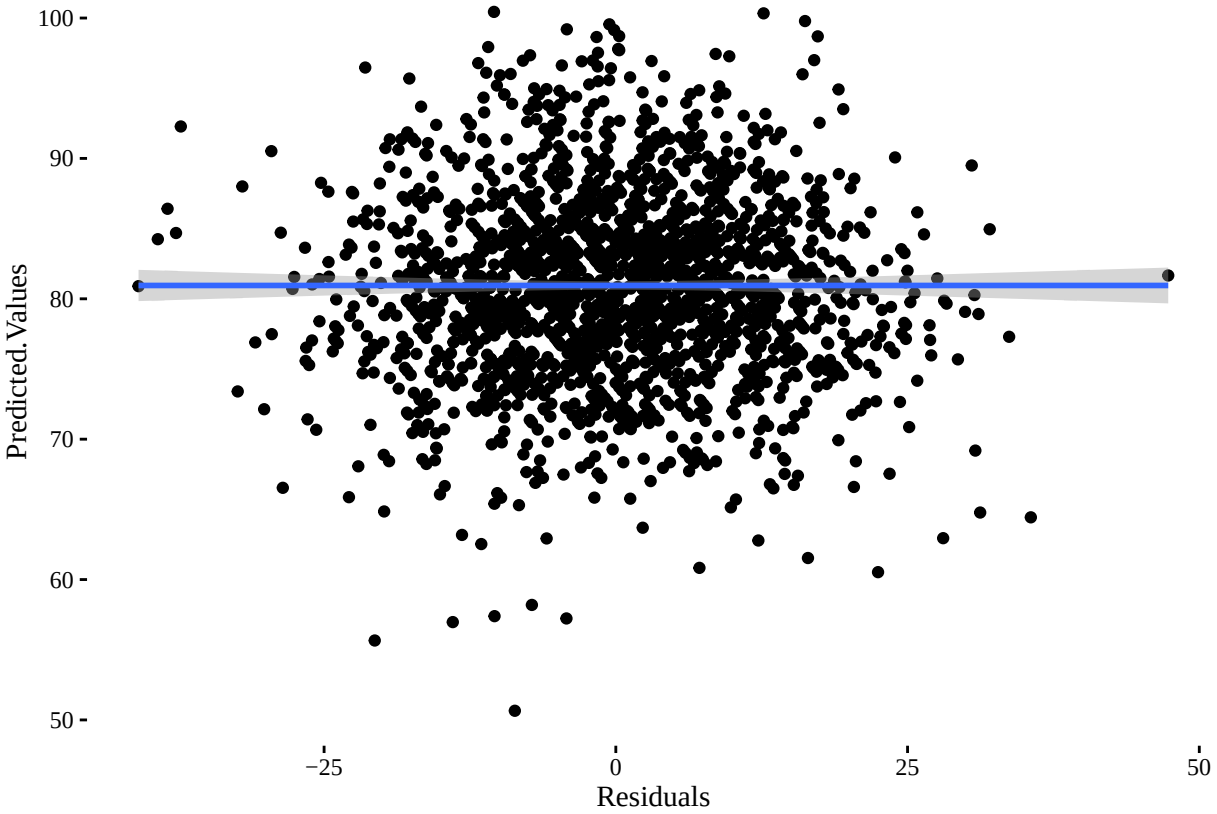
In the above model, errors is the only variable that has been transformed so far. I took its natural log. All of our variables are highly significant, and we have a very low p-value on our f-statistic. This does suggest we could add more variables from our dataset, but there are a lot of conceptual and observed relationships between these variables. This model, while sparse, appropriately brings in enough unique information to be parsimonious.

Now that I've arrived at my model choice, I'll check a box-cox plot to see if I should perform a transformation on my dependent variable:



While the 95% confidence interval doesn't include $\lambda = 1$, It's entirely between 1 and 1.5. I think in this case, rather than trying out a few values between 1 and 1.5, I'll stick with no box cox transformation, assuming 1 is close enough.

Next, lets check the residuals versus the fitted values.



This scatter plot of our predicted values versus residuals seems to suggest no underlying relationship.

Conclusions

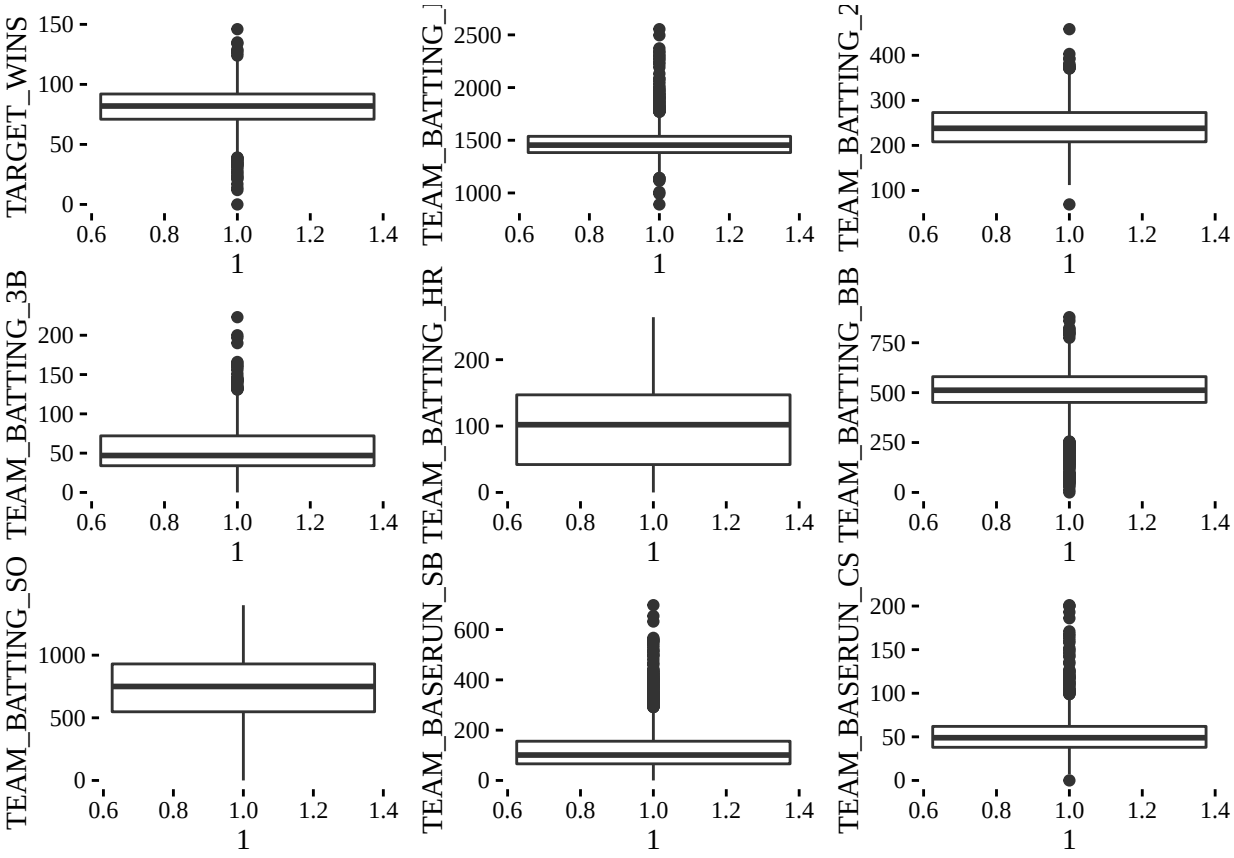
This model would benefit from more conceptual analysis of what these measures mean, and more analysis of how this compares to the observed relationships between the variables.

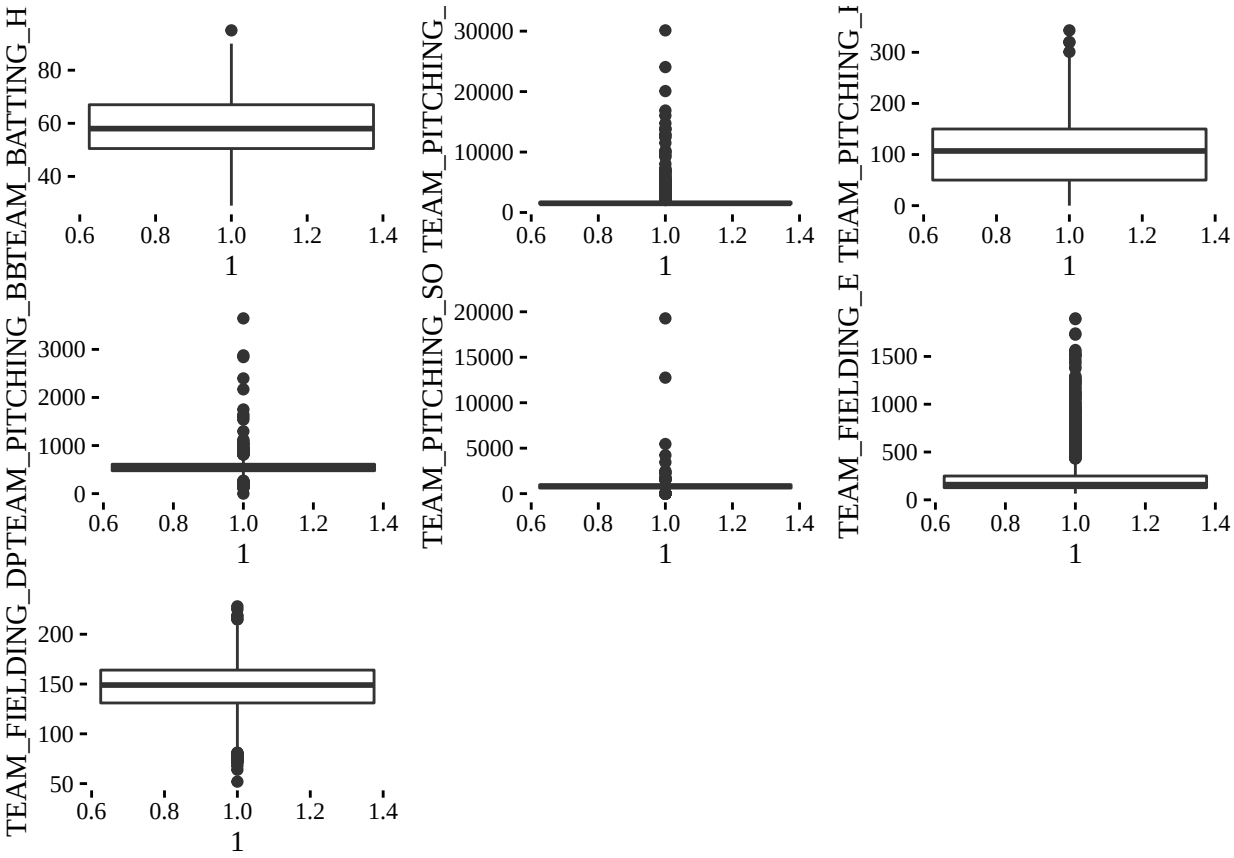
It's pretty clear that there should be relationships betweeen base hits, singles, doubles, triples, and home runs, but looking at pairwise correlation, only some pairs showed correlation. There's no clear reason why base hits should be correlated to doubles, but not triples or home runs for example.

I'd also like to dig into the definition of errors, and research more about the relationships between measures that appeared in both pitcher and batter. It would appear some of these should be negatively correlated, but I'd be interested to learn more about what sorts of differences might be present in these measures. If they're exactly 1-1, I could try to eliminate one side or the other.

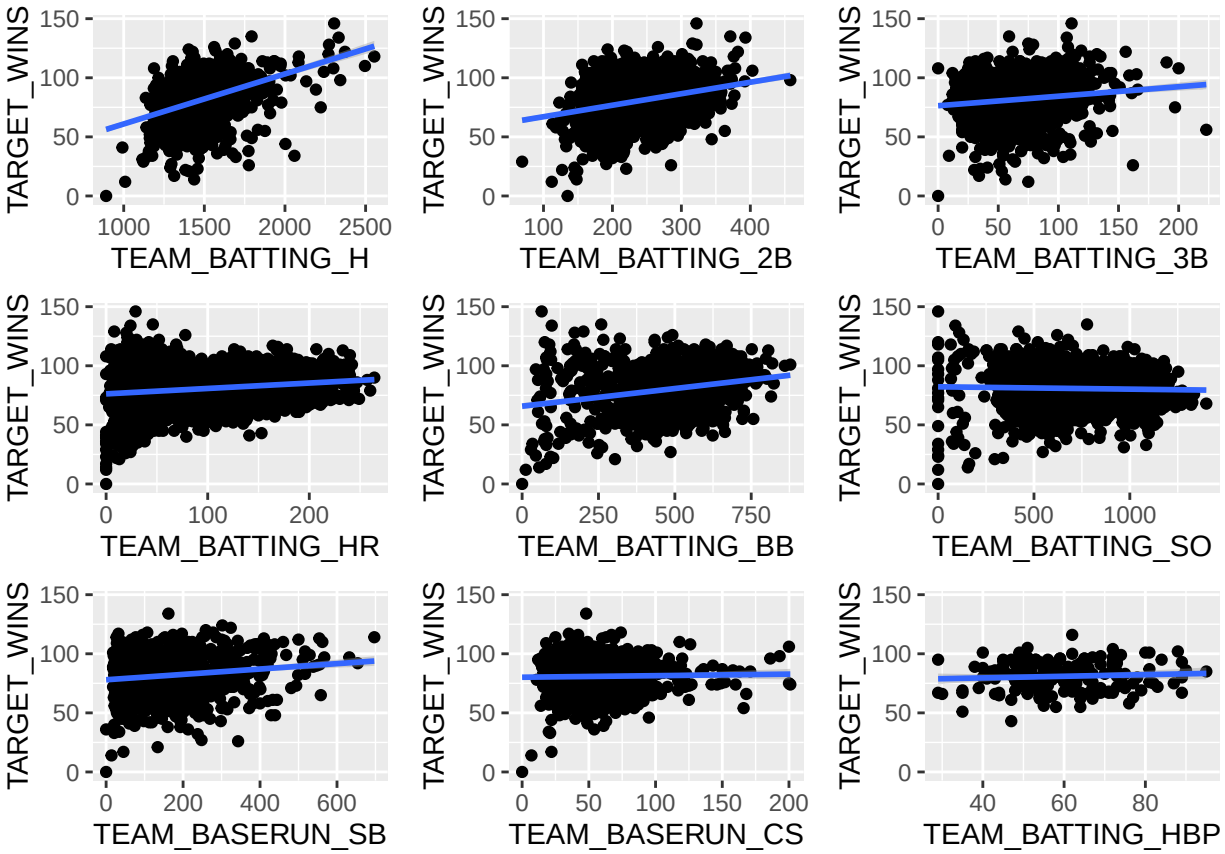
Appendex

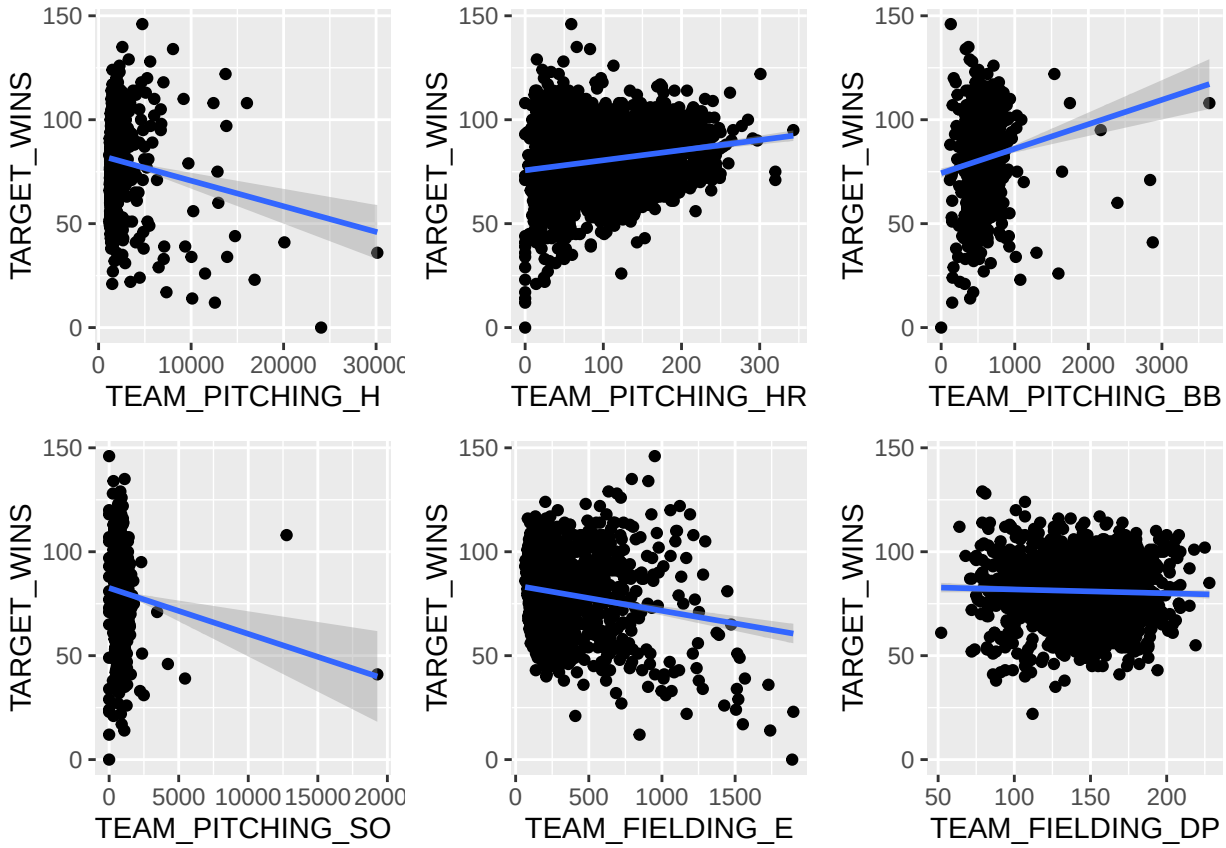
Boxplots of original data



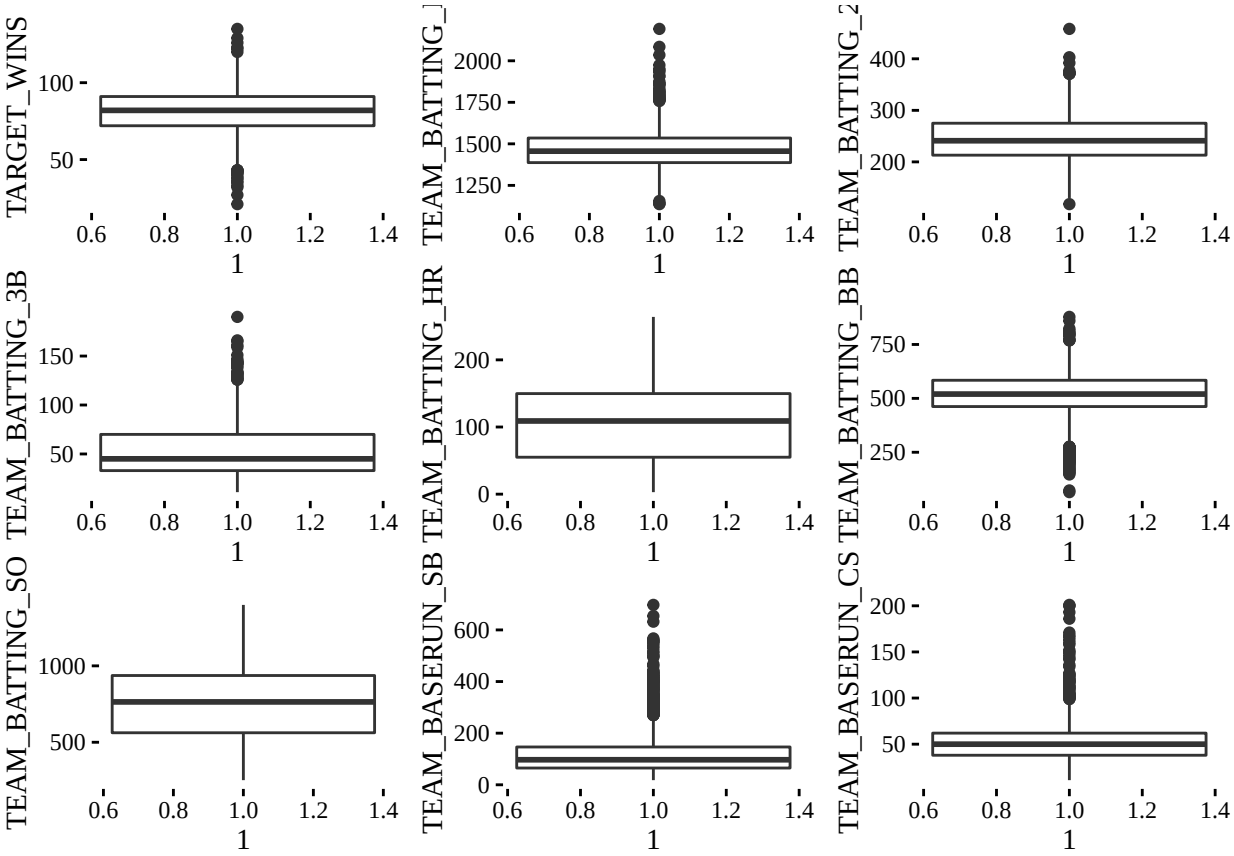

TEAM_BATTING_H
80
60
40
0.6
0.8
1.0
1.2
1.4
1
TEAM_PITCHING_BB
3000
2000
1000
0
TEAM_PITCHING_SO
30000
20000
10000
15000
5000
TEAM_PITCHING_H
300
200
100
TEAM_FIELDING_DP
150
50
TEAM_FIELDING_E
1500
1000
500

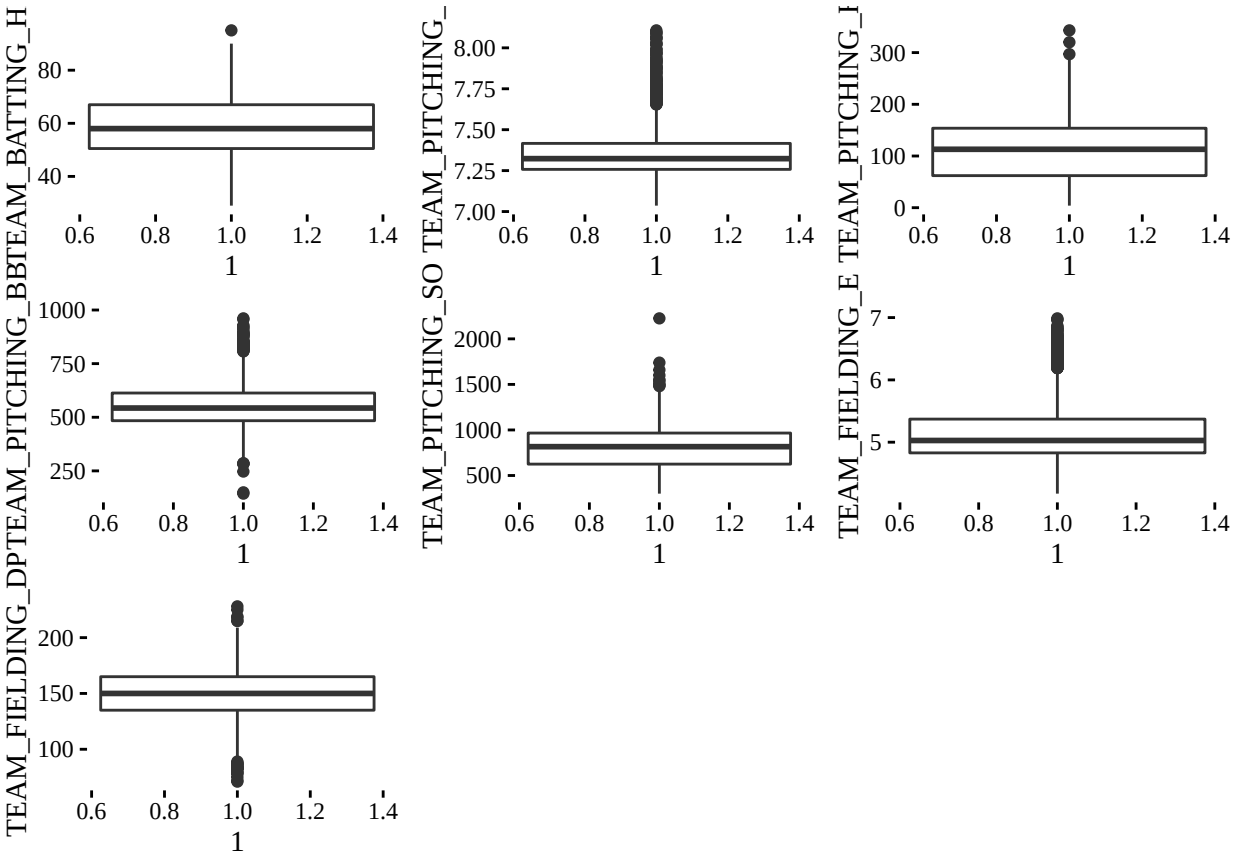
Scatterplots of original data: variable vs WINS



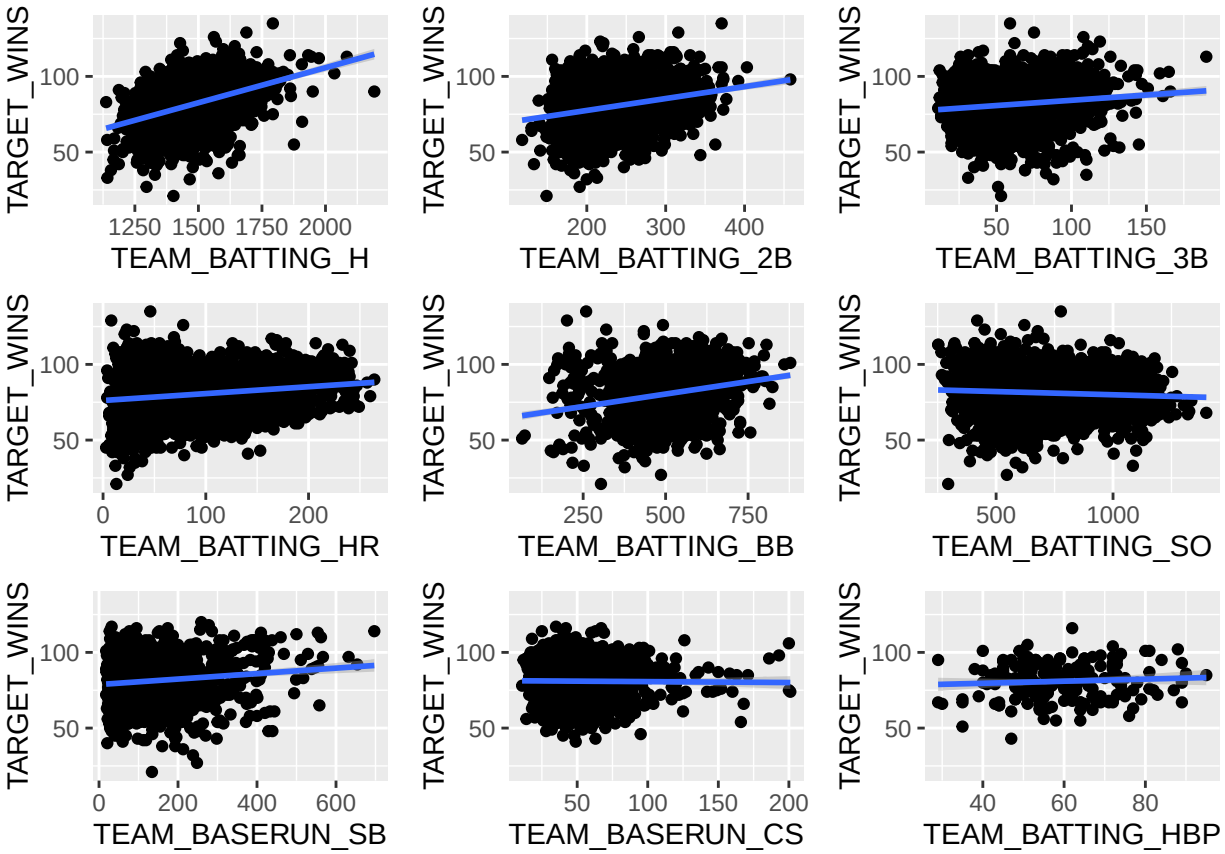

TARGET_WINS
TEAM_PITCHING_H
TEAM_PITCHING_HR
TEAM_PITCHING_BB
TEAM_PITCHING_SO
TEAM_FIELDING_E
TEAM_FIELDING_DP

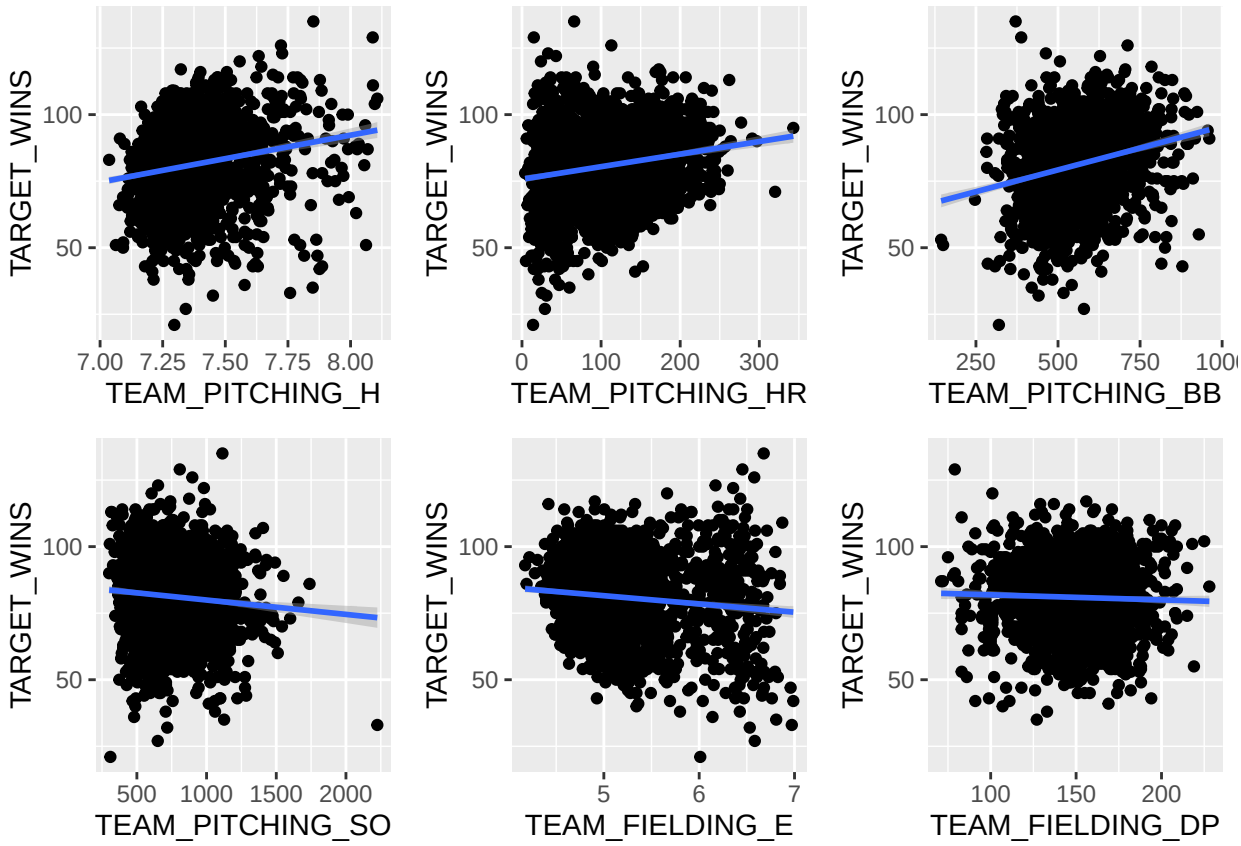
Boxplots of data with transformations applied




TEAM_BATTING_H
TEAM_BATTING_BB
TEAM_PITCHING_H
TEAM_PITCHING_BB
TEAM_PITCHING_SO
TEAM_FIELDING_E
TEAM_FIELDING_DP
1

Scatterplots of data with transformations applied, variable vs WINS




TARGET_WINS
100
50
7.00 7.25 7.50 7.75 8.00
TEAM_PITCHING_H
TARGET_WINS
100
50
0 100 200 300
TEAM_PITCHING_HR
TARGET_WINS
100
50
250 500 750
TEAM_PITCHING_BB
TARGET_WINS
100
50
500 1000 1500 2000
TEAM_PITCHING_SO
TARGET_WINS
100
50
5 6 7
TEAM_FIELDING_E
TARGET_WINS
100
50
100 150 200
TEAM_FIELDING_DP