Moneyball

Biguzzi, Connin, Greenlee, Moscoe, Sooklall, Telab, and Wright

9/23/2021

## Introduction

How can we predict the number of wins for a baseball team in a given season based on information about its performance during the season? In this report, we construct a linear model to answer this question. While our model is based on observational data rather than a randomized controlled experiment, it may lend some insight into what elements of a baseball team’s performance are most important in securing wins. Variables that are strongly associated with a greater number of wins for a team would be worthy of further study.

We begin this report by describing the available data. The second section includes our transformation of the data so that it conforms as closely as possible to the conditions for linear modeling. Then we compare several models. Finally, we select an optimal model and use it to generate predictions for an evaluation data set.

## Data Exploration

Exploring the data is essential to uncovering any surprises or unusual features in the data. As we explore the data, we also bear in mind the conditions for linear modeling:

* The residuals of the model should be nearly normal;
* The variability of the residuals should be nearly constant;
* The residuals are independent;
* Each variable is linearly related to the target variable.

While almost no real-world data set conforms perfectly to these conditions, there may be transformations that we can perform on the data to alleviate some of its violations of these conditions.

The data set contains 2,276 rows each representing the full-season performance of a baseball team between the years 1871 and 2006 inclusive.

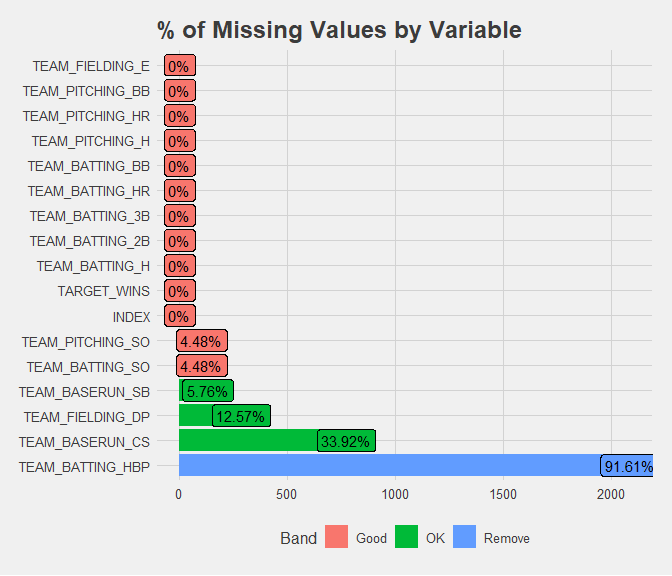
Let’s look first at a numerical summary of each variable:

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| INDEX | 0 | 1 | 1268 | 736 | 1 | 631 | 1270 | 1916 | 2535 |
| TARGET\_WINS | 0 | 1 | 81 | 16 | 0 | 71 | 82 | 92 | 146 |
| TEAM\_BATTING\_H | 0 | 1 | 1469 | 145 | 891 | 1383 | 1454 | 1537 | 2554 |
| TEAM\_BATTING\_2B | 0 | 1 | 241 | 47 | 69 | 208 | 238 | 273 | 458 |
| TEAM\_BATTING\_3B | 0 | 1 | 55 | 28 | 0 | 34 | 47 | 72 | 223 |
| TEAM\_BATTING\_HR | 0 | 1 | 100 | 61 | 0 | 42 | 102 | 147 | 264 |
| TEAM\_BATTING\_BB | 0 | 1 | 502 | 123 | 0 | 451 | 512 | 580 | 878 |
| TEAM\_BATTING\_SO | 102 | 1 | 736 | 249 | 0 | 548 | 750 | 930 | 1399 |
| TEAM\_BASERUN\_SB | 131 | 1 | 125 | 88 | 0 | 66 | 101 | 156 | 697 |
| TEAM\_BASERUN\_CS | 772 | 1 | 53 | 23 | 0 | 38 | 49 | 62 | 201 |
| TEAM\_BATTING\_HBP | 2085 | 0 | 59 | 13 | 29 | 50 | 58 | 67 | 95 |
| TEAM\_PITCHING\_H | 0 | 1 | 1779 | 1407 | 1137 | 1419 | 1518 | 1682 | 30132 |
| TEAM\_PITCHING\_HR | 0 | 1 | 106 | 61 | 0 | 50 | 107 | 150 | 343 |
| TEAM\_PITCHING\_BB | 0 | 1 | 553 | 166 | 0 | 476 | 536 | 611 | 3645 |
| TEAM\_PITCHING\_SO | 102 | 1 | 818 | 553 | 0 | 615 | 814 | 968 | 19278 |
| TEAM\_FIELDING\_E | 0 | 1 | 246 | 228 | 65 | 127 | 159 | 249 | 1898 |
| TEAM\_FIELDING\_DP | 286 | 1 | 146 | 26 | 52 | 131 | 149 | 164 | 228 |

From the summary above, we can see that some variables show a significant number of missing values, especially TEAM\_BATTING\_HBP and TEAM\_BASERUN\_CS. Some variables show suspiciously large maximum values, such as TEAM\_PITCHING\_H, TEAM\_PITCHING\_BB, TEAM\_PITCHING\_SO, and TEAM\_FIELDING\_E. We can also see that some variables contain entries of zero that don’t make sense in the context of a baseball season. These variables include TARGET\_WINS, TEAM\_BATTING\_3B, TEAM\_BATTING\_HR, TEAM\_BATTING\_BB, TEAM\_BATTING\_SO, TEAM\_BASERUN\_SB, TEAM\_BASERUN\_CS, TEAM\_PITCHING\_HR, TEAM\_PITCHING\_BB, TEAM\_PITCHING\_SO. At least some of these entries we know to be erroneous. For example, the [all-time minimum](https://www.baseball-almanac.com/recbooks/rb_strike2.shtml) batting strikeouts for a team over a complete season was 308, achieved by the 1921 Cincinnati Reds.

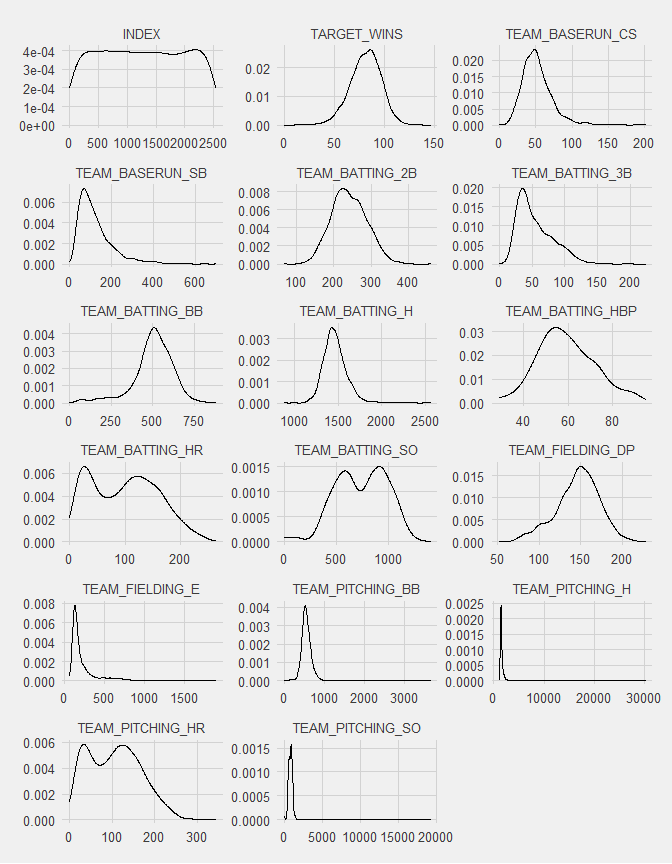
How we respond to the missingness in the data will depend on how it is distributed across columns and rows.



Here we see again that missing entries are not distributed randomly throughout the data set. In particular, TEAM\_BATTING\_HBP is comprised of almost 92% missing values. This variable cannot provide much information to our model and will be dropped. TEAM\_BASERUN\_CS also displays a very high fraction of missing values: 34%. TEAM\_PITCHING\_SO and TEAM\_BATTING\_SO share the exact same number of missing values. This suggests that the missingness in these variables is not random.

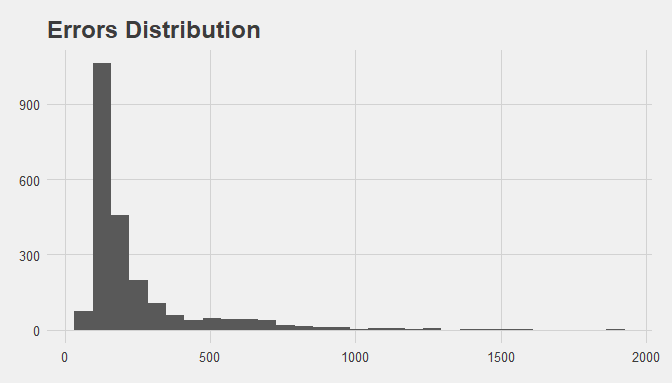
Further investigation reveals that the distribution of missingness in both these variables is identical. One option for responding to the missingness in these variables will be to code it as its own factor variable. Another option may be to drop one or both these variables from our models if there are further reasons to be suspicious about the quality of the data contained in them.

Examining the shape of each variable can help identify any unusual attributes including extreme skewness or low variance.

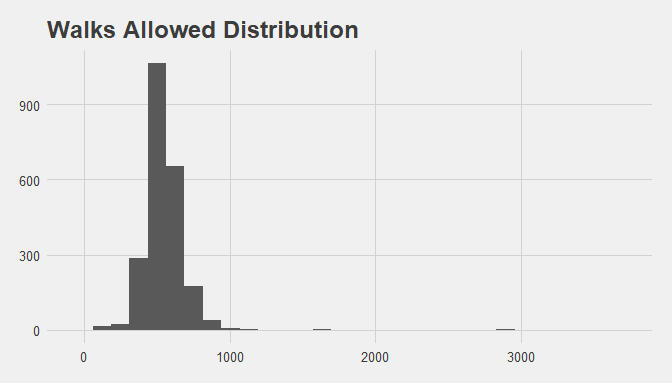


Density plots for each variable reveals the highly skewed shapes of TEAM\_FIELDING\_E, TEAM\_PITCHING\_BB, TEAM\_PITCHING\_H, and TEAM\_PITCHING\_SO as suggested by the numeric summary above. TEAM\_BASERUN\_SB and TEAM\_BATTING\_3B display moderate skewness. Other distributions appear roughly normal or bimodal.

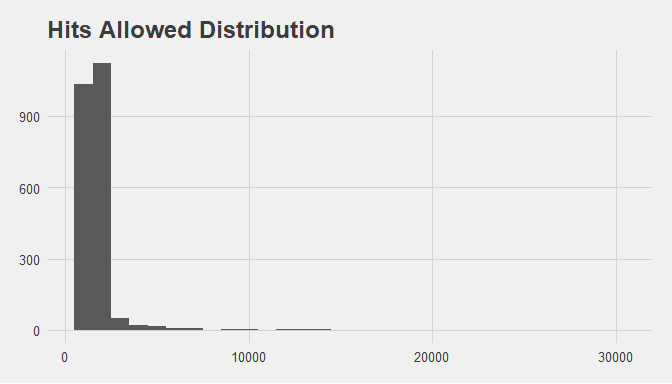
Looking more closely at each of the highly skewed variables may suggest a reason for their unexpected shape.



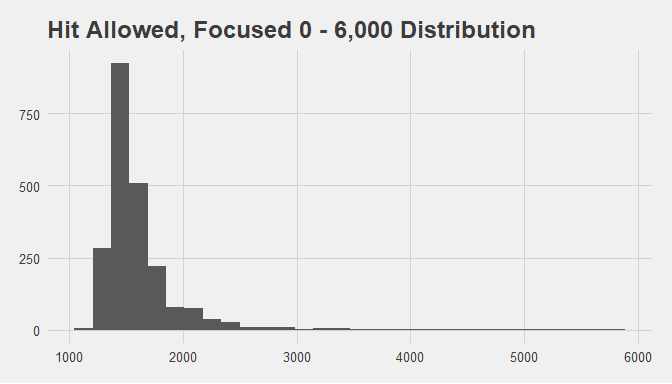
In TEAM\_FIELDING\_E, a very small number of entries exceed 1000. Excluding these, the distribution is moderately skewed right.



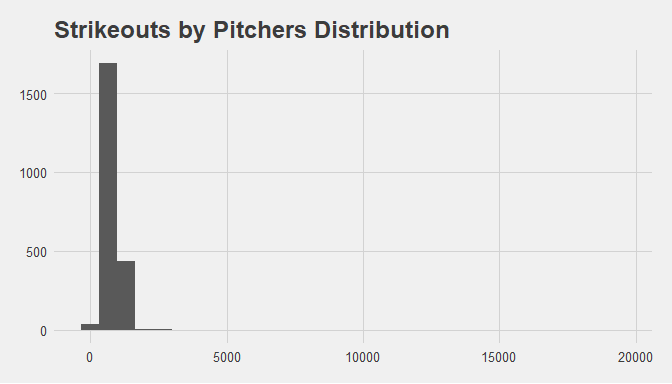
The distribution of TEAM\_PITCHING\_BB appears normal except for a very small number of entries greater than 1000.



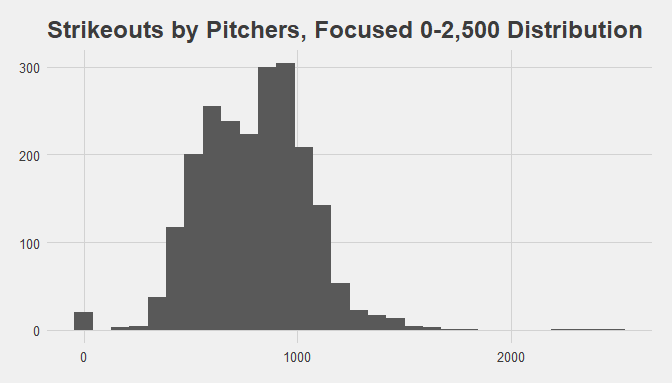
Let’s look more closely at the region where most of the values in TEAM\_PITCHING\_H lie, [0,6000]:



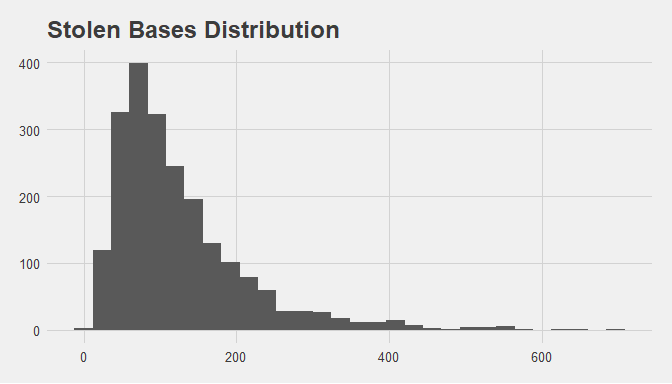
TEAM\_PITCHING\_H is skewed even within this narrower region surrounding the peak. This suggests possible data entry error for values greater than 3250.



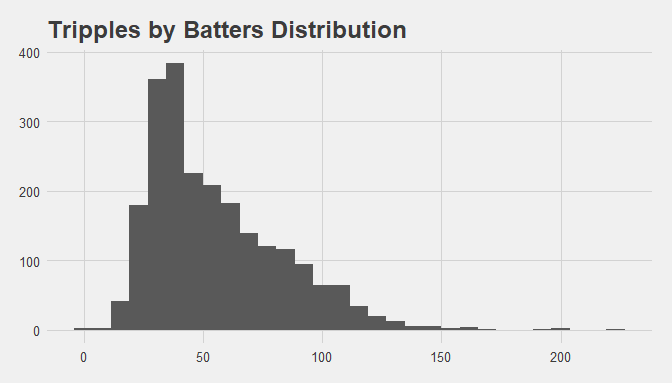
Narrowing in on TEAM\_PITCHING\_SO:



The data appear roughly normal, except for a significant number of unrealistic zero-values. This suggests possible data entry error for values of zero, or greater than or equal to 2000.



The shape of TEAM\_BASERUN\_SB does not suggest any data entry errors, even though there exist a few extreme outliers.

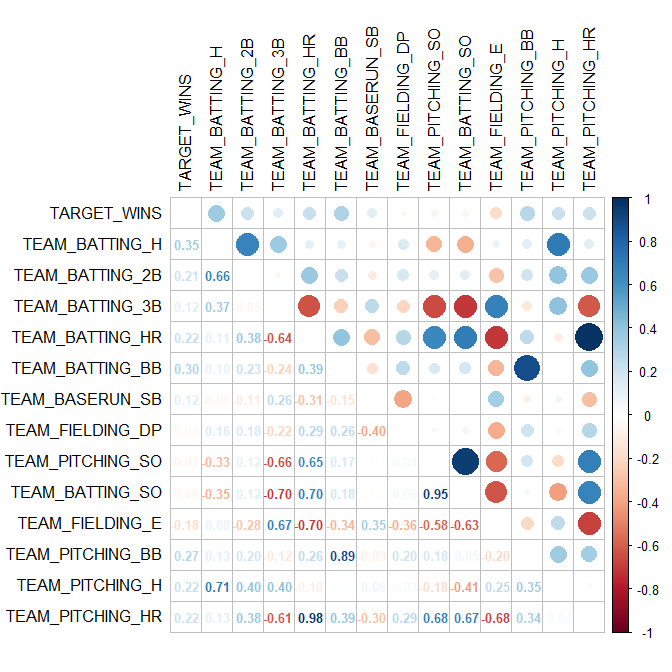


The shape of TEAM\_BATTING\_3B does not suggest any data entry errors, even though there exist a few extreme outliers.

At this point in our data exploration, we’re aware that several variables contain outliers so extreme that their accuracy is questionable. We also know for sure that at least some data is erroneous. The best course of action under these conditions is to pause our investigation and look for ways to improve the quality of the data. Because this is not possible, we’ll use multiple imputation to replace extreme outliers and missing values. This decision entails some risk, because replacing data with imputed values could impair the quality of our model. However in this case, it’s our judgment that incorporating these extreme, unrealistic values would do greater damage to any models that are trained on them. In order to proceed with caution, we’ll train models on two versions of the data–one with extreme values replaced by imputation, and one with extreme values unchanged.

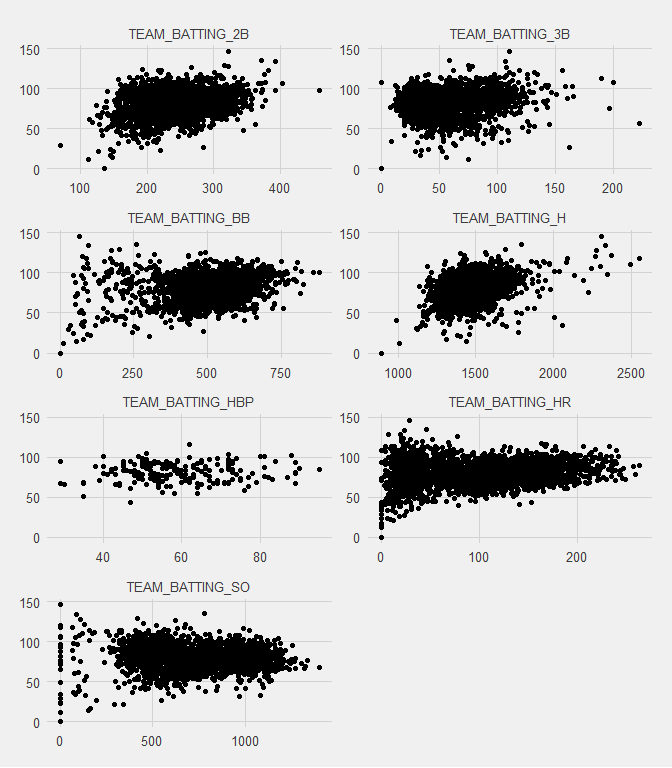
One important condition of linear modeling is that independent variables should be uncorrelated with each other. In real-world data it is rarely possible to fully satisfy this condition. However, we can seek to avoid large pairwise correlations between variables.

The correlation plot below is arranged so that all variables with a theoretical positive effect on TARGET\_WINS appear at the top and left, followed by variables with a theoretical negative effect. We would expect positive correlations among variables with positive effects, positive correlations among variables with negative effects, and negative correlations among variables with opposite effects.

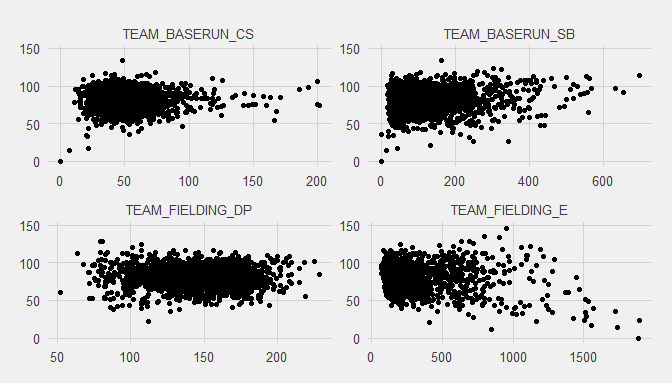


But this is not what we see. Instead we see that no variable alone has correlation greater than 0.35 in either direction with the target. We also see that some pairs of variables have correlations that are so strong or misdirected that we have reason to doubt the integrity of the data. Note that TEAM\_PITCHING\_HR and TEAM\_BATTING\_HR, with correlation 0.98, contain essentially the same information. TEAM\_PITCHING\_SO and TEAM\_BATTING\_SO, the variables whose patterns of missingness are identical, also demonstrate an unreasonably high correlation of 0.95.

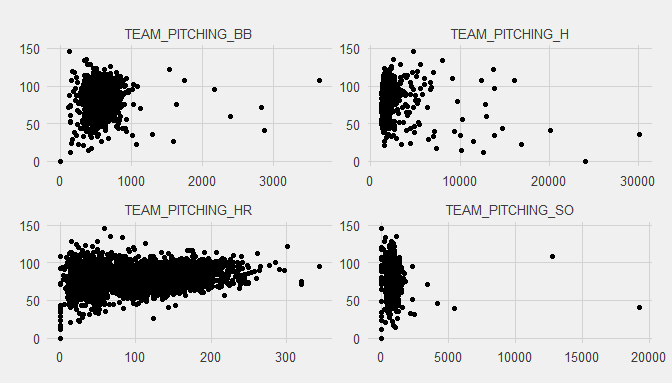
Another condition of linear models is that independent variables be linearly related to the target. We can examine this using scatterplots, one group of variables at a time.

*Batting variables* 

No batting variable exhibits a strong positive or negative linear relationship with the target, although the scatterplots also show that neither does any batting variable exhibit a strong non-linear relationship with TARGET\_WINS.

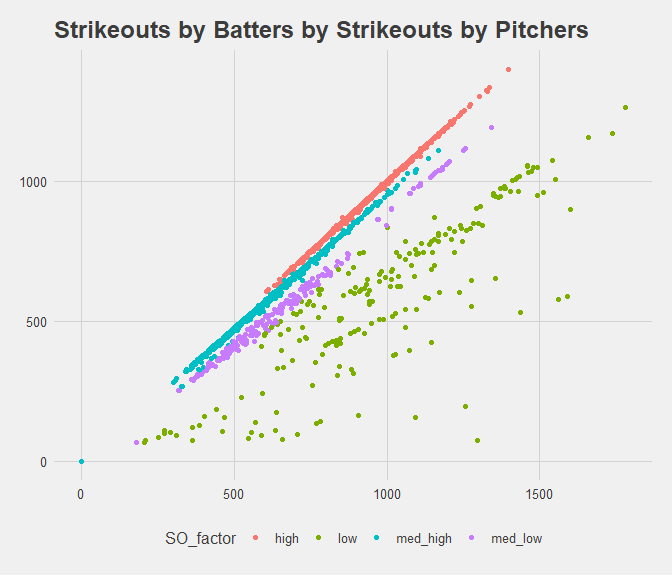
*Baserun and fielding variables* 

The situation with these variables is similar to that of the batting variables.

*Pitching variables* 

No nonlinearity is evident in the scatterplots for the pitching variables. What is evident again is the presence of extreme outliers and skewness in three of the four pitching variables.

An unexpected finding in our exploratory data analysis is that a scatterplot of TEAM\_BATTING\_SO and TEAM\_PITCHING\_SO suggests that the data can be divided into 4 distinct groups, three of which contain highly correlated data.



There is no theoretical reason to expect such a grouping, or to expect such high correlations between the number of strikeouts a team incurs while batting, and the number of strikeouts a team achieves while pitching. However, it may be useful to divide the data into the four groups suggested by these relationships for modeling purposes.

## Data Preparation

To prepare the data for modeling, we first simplify variable names by removing the prefix TEAM\_, and we add the grouping variable SO\_FACTOR.

An understanding of baseball suggests how to combine some variables into other meaningful measures of team performance. Creating new variables as combinations of existing variables will help alleviate collinearity among some pairs of variables, and it will lead to models that better reflect commonly used measures of team performance. We create variables to represent net stolen bases (BASERUN\_NET\_SB), offensive on-base percentage (OFFENSE\_OBP), defensive on-base percentage (DEFENSE\_OBP), and total at-bats (TOT\_AT\_BATS). Then we drop the variables that were combined to produce these new variables.

After these initial transformations, we split the data into a training set (80% = 1820 rows) and a test set (20% = 456 rows).

During our exploration of the data, we noted several variables with extreme outliers and described our strategy of examining two versions of our models: one trained on data with only missing values imputed, and one trained on data with missing values and outliers imputed. We’ll consider all these models before making our final selection.

A look at the summaries of our test sets, both with and without outliers imputed, shows the raw materials for our models.

*With missing values imputed*

**Variable type: character**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| SO\_FACTOR | 0 | 1 | 3 | 8 | 0 | 4 | 0 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| INDEX | 0 | 1 | 1266 | 739 | 1 | 621 | 1280 | 1903 | 2534 |
| BATTING\_2B | 0 | 1 | 242 | 47 | 69 | 209 | 238 | 274 | 458 |
| BATTING\_3B | 0 | 1 | 55 | 28 | 0 | 34 | 47 | 72 | 223 |
| BATTING\_HR | 0 | 1 | 101 | 61 | 0 | 44 | 104 | 148 | 264 |
| BATTING\_SO | 0 | 1 | 731 | 253 | 0 | 543 | 748 | 928 | 1399 |
| PITCHING\_HR | 0 | 1 | 107 | 61 | 0 | 53 | 108 | 152 | 343 |
| PITCHING\_SO | 0 | 1 | 810 | 592 | -248 | 603 | 810 | 966 | 19278 |
| BASERUN\_NET\_SB | 0 | 1 | 38 | 40 | -140 | 13 | 35 | 60 | 628 |
| OFFENSE\_OBP | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DEFENSE\_OBP | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| TOT\_AT\_BATS | 0 | 1 | 6325 | 226 | 1270 | 6217 | 6317 | 6424 | 7518 |
| TARGET\_WINS | 0 | 1 | 81 | 16 | 0 | 71 | 82 | 92 | 146 |

*With missing values and outliers imputed*

**Variable type: character**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| SO\_FACTOR | 0 | 1 | 3 | 8 | 0 | 4 | 0 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| INDEX | 0 | 1 | 1266 | 739 | 1 | 621 | 1280 | 1903 | 2534 |
| BATTING\_2B | 0 | 1 | 241 | 41 | 120 | 211 | 238 | 272 | 330 |
| BATTING\_3B | 0 | 1 | 54 | 24 | 3 | 35 | 47 | 70 | 117 |
| BATTING\_HR | 0 | 1 | 100 | 57 | -35 | 46 | 104 | 147 | 232 |
| BATTING\_SO | 0 | 1 | 739 | 233 | -788 | 556 | 748 | 925 | 1246 |
| PITCHING\_HR | 0 | 1 | 106 | 57 | -2 | 56 | 109 | 150 | 243 |
| PITCHING\_SO | 0 | 1 | 798 | 220 | 289 | 620 | 810 | 957 | 1453 |
| BASERUN\_NET\_SB | 0 | 1 | 38 | 30 | -54 | 16 | 36 | 57 | 129 |
| OFFENSE\_OBP | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DEFENSE\_OBP | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| TOT\_AT\_BATS | 0 | 1 | 6314 | 131 | 5898 | 6219 | 6309 | 6401 | 6723 |
| TARGET\_WINS | 0 | 1 | 81 | 16 | 0 | 71 | 82 | 92 | 146 |

## Build Models

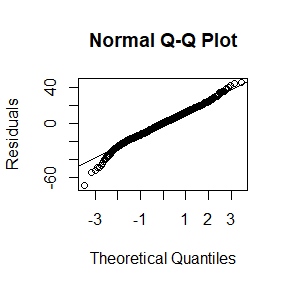
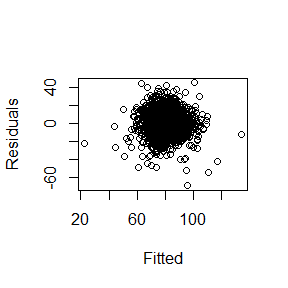
In this section we construct models first for training sets with only missing values imputed, and we evaluate them with a test set with only missing values imputed. Then, we train and test analogous models with outliers imputed as well.

### Model 1. Almost all variables

This model contains all variables except BATTING\_HR, BATTING\_SO, and TOT\_AT\_BATS, because each of these variables is highly correlated with at least one other variable.

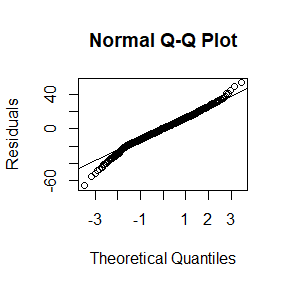
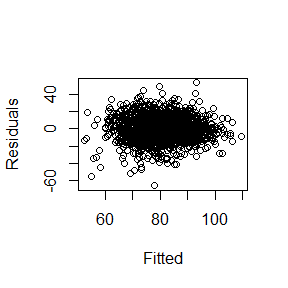
*With only missing values imputed*

##   
## Call:  
## lm(formula = TARGET\_WINS ~ BATTING\_2B + BATTING\_3B + PITCHING\_HR +   
## PITCHING\_SO + SO\_FACTOR + BASERUN\_NET\_SB + OFFENSE\_OBP +   
## DEFENSE\_OBP, data = train\_imp\_M)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -69.823 -8.680 0.287 8.508 45.451   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.896e+01 6.597e+00 -2.873 0.004108 \*\*   
## BATTING\_2B -6.953e-04 8.238e-03 -0.084 0.932743   
## BATTING\_3B 1.759e-01 1.637e-02 10.746 < 2e-16 \*\*\*  
## PITCHING\_HR 8.232e-02 8.225e-03 10.009 < 2e-16 \*\*\*  
## PITCHING\_SO -4.838e-03 2.216e-03 -2.184 0.029103 \*   
## SO\_FACTORlow -9.508e+00 1.341e+00 -7.089 1.93e-12 \*\*\*  
## SO\_FACTORmed\_high -6.993e-01 8.868e-01 -0.789 0.430476   
## SO\_FACTORmed\_low -5.064e+00 1.324e+00 -3.825 0.000135 \*\*\*  
## BASERUN\_NET\_SB 1.301e-01 9.733e-03 13.368 < 2e-16 \*\*\*  
## OFFENSE\_OBP 2.202e+02 2.358e+01 9.338 < 2e-16 \*\*\*  
## DEFENSE\_OBP 3.620e+01 8.948e+00 4.045 5.45e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.95 on 1809 degrees of freedom  
## Multiple R-squared: 0.3183, Adjusted R-squared: 0.3146   
## F-statistic: 84.48 on 10 and 1809 DF, p-value: < 2.2e-16



*With missing values and outliers imputed*

##   
## Call:  
## lm(formula = TARGET\_WINS ~ BATTING\_2B + BATTING\_3B + PITCHING\_HR +   
## PITCHING\_SO + SO\_FACTOR + BASERUN\_NET\_SB + OFFENSE\_OBP +   
## DEFENSE\_OBP, data = train\_imp\_OM)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -65.998 -8.317 0.393 8.345 53.015   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -66.553003 8.956207 -7.431 1.65e-13 \*\*\*  
## BATTING\_2B -0.043542 0.009733 -4.474 8.16e-06 \*\*\*  
## BATTING\_3B 0.208312 0.020901 9.967 < 2e-16 \*\*\*  
## PITCHING\_HR 0.083258 0.008694 9.576 < 2e-16 \*\*\*  
## PITCHING\_SO -0.007421 0.002402 -3.089 0.00204 \*\*   
## SO\_FACTORlow 0.364991 1.820872 0.200 0.84115   
## SO\_FACTORmed\_high -3.629889 0.923957 -3.929 8.86e-05 \*\*\*  
## SO\_FACTORmed\_low -5.804593 1.383176 -4.197 2.84e-05 \*\*\*  
## BASERUN\_NET\_SB 0.117435 0.011785 9.965 < 2e-16 \*\*\*  
## OFFENSE\_OBP 487.923479 34.888081 13.985 < 2e-16 \*\*\*  
## DEFENSE\_OBP -40.208228 17.622460 -2.282 0.02263 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.1 on 1809 degrees of freedom  
## Multiple R-squared: 0.3023, Adjusted R-squared: 0.2984   
## F-statistic: 78.38 on 10 and 1809 DF, p-value: < 2.2e-16



*Evaluation and discussion*

The root mean square error for each model is reported below.

## [1] "Test set RMSE for model 1 with only missing imputed:"

## [1] 15.56565

## [1] "Test set RMSE for model 1 with missing and outliers imputed:"

## [1] 14.58961

Our initial look at the residuals for the model with only missing values imputed shows some heteroskedasticity, violating one of our modeling conditions. It appears the variance of the residuals decreases at extreme values. We also notice from the QQ plot that the residuals depart from a normal distribution at the extremes.

Plots for the model trained on the imputed outliers show improvement on the issues of the previous model. This could be a hint that the extreme outliers that have been removed were affecting our model.

RMSE values for these models on the test set are similar to the RMSE values for the training sets. This means that these models avoid overfitting the data.

### Model 2. Piecewise by FACTOR\_SO

The scatterplot of BATTING\_SO vs. PITCHING\_SO suggested four groups in this data. Here we fit one model to each group. Since we conduct a piecewise analysis on two separate training sets, 8 sets of outputs are generated for models in this section. Rather than display this model output, we summarize our findings.

*Evaluation and discussion*

The root mean square error for each model is reported below.

## [1] "RMSE on training set for piecewise model with missing values imputed:"

## [1] 12.29754

## [1] "RMSE on test set for piecewise model with missing values imputed:"

## [1] 15.42546

## [1] "RMSE on training set for piecewise model with missing values and outliers imputed:"

## [1] 12.66628

## [1] "RMSE on test set for piecewise model with missing values and outliers imputed:"

## [1] 14.82223

This piecewise modeling approach returned RMSE values similar to those of our previous model. Residual and QQ plots for component models where only missing values were imputed demonstrate the same heteroskedasticity and deviation from normality that we saw in Model 1. Again, these issues are somewhat mitigated in the model where outliers are also imputed.

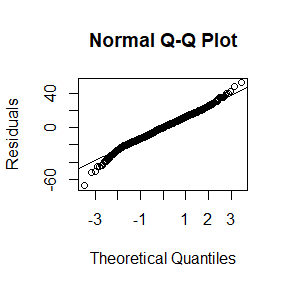
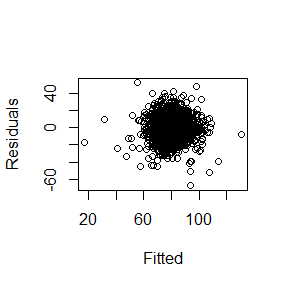
This second model was prompted by the results of our BATTING\_SO vs. PITCHING\_SO plot. We acknowledge some issues with this approach– one being that the models differ greatly in population size, so a visual comparison may not be trustworthy. Also, we may not have the appropriate context to make the decision to divide the data based on the relationship between these two predictors. While this might have led to overfitting, RMSE values that are similar in both the training and test sets suggest that we avoided this.

### Model 3. Drop DEFENSE\_OBP

Here we drop DEFENSE\_OBP from Model 1. Even though this variable is a statistically significant predictor of TARGET\_WINS, its large coefficient and standard error compared to the other variables might be due to its high correlation with OFFENSE\_OBP.

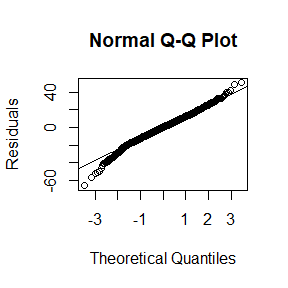
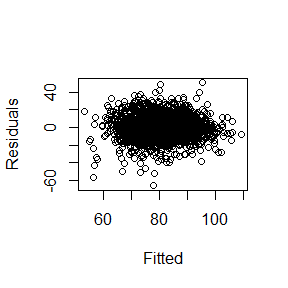
*With only missing values imputed*

##   
## Call:  
## lm(formula = TARGET\_WINS ~ BATTING\_2B + BATTING\_3B + PITCHING\_HR +   
## PITCHING\_SO + SO\_FACTOR + BASERUN\_NET\_SB + OFFENSE\_OBP, data = train\_imp\_M)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -67.739 -8.813 0.420 8.388 52.228   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.886e+01 6.152e+00 -4.691 2.93e-06 \*\*\*  
## BATTING\_2B 5.500e-04 8.267e-03 0.067 0.946969   
## BATTING\_3B 1.925e-01 1.591e-02 12.096 < 2e-16 \*\*\*  
## PITCHING\_HR 6.533e-02 7.102e-03 9.200 < 2e-16 \*\*\*  
## PITCHING\_SO 3.299e-03 9.323e-04 3.539 0.000412 \*\*\*  
## SO\_FACTORlow -7.935e+00 1.289e+00 -6.155 9.20e-10 \*\*\*  
## SO\_FACTORmed\_high 7.598e-01 8.135e-01 0.934 0.350463   
## SO\_FACTORmed\_low -3.925e+00 1.299e+00 -3.021 0.002555 \*\*   
## BASERUN\_NET\_SB 1.230e-01 9.612e-03 12.794 < 2e-16 \*\*\*  
## OFFENSE\_OBP 2.695e+02 2.029e+01 13.281 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.01 on 1810 degrees of freedom  
## Multiple R-squared: 0.3122, Adjusted R-squared: 0.3087   
## F-statistic: 91.27 on 9 and 1810 DF, p-value: < 2.2e-16



*With missing values and outliers imputed*

##   
## Call:  
## lm(formula = TARGET\_WINS ~ BATTING\_2B + BATTING\_3B + PITCHING\_HR +   
## PITCHING\_SO + SO\_FACTOR + BASERUN\_NET\_SB + OFFENSE\_OBP, data = train\_imp\_OM)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -65.971 -8.208 0.427 8.421 50.449   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -66.136108 8.964740 -7.377 2.45e-13 \*\*\*  
## BATTING\_2B -0.042919 0.009740 -4.406 1.11e-05 \*\*\*  
## BATTING\_3B 0.195525 0.020159 9.699 < 2e-16 \*\*\*  
## PITCHING\_HR 0.082616 0.008700 9.497 < 2e-16 \*\*\*  
## PITCHING\_SO -0.007986 0.002392 -3.339 0.000859 \*\*\*  
## SO\_FACTORlow -2.657825 1.250600 -2.125 0.033702 \*   
## SO\_FACTORmed\_high -4.107337 0.900994 -4.559 5.49e-06 \*\*\*  
## SO\_FACTORmed\_low -6.804550 1.313431 -5.181 2.46e-07 \*\*\*  
## BASERUN\_NET\_SB 0.119766 0.011754 10.189 < 2e-16 \*\*\*  
## OFFENSE\_OBP 449.443838 30.576500 14.699 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.12 on 1810 degrees of freedom  
## Multiple R-squared: 0.3003, Adjusted R-squared: 0.2968   
## F-statistic: 86.31 on 9 and 1810 DF, p-value: < 2.2e-16



*Evaluation and discussion*

The root mean square error for each model is reported below.

## [1] "RMSE on training set for model 3 with missing values imputed:"

## [1] 15.41436

## [1] "RMSE on training set for model 3 with missing values and outliers imputed:"

## [1] 14.59399

In this model, we see residuals behaving similarly to the first model. We can call attention again to the improvement in the residuals by the training of the data with the imputed outliers. Not only is the shape of the residuals more homoskedastic and normal, but the statistical significance of the coefficients is much greater.

## Conclusion and model selection

Before we select the best performing model, let’s review the above analysis.

*Data Exploration*. We recognize a high probability of flawed data. Unrelated variables were found to have near perfect correlation. Outliers represent impossible measurements. These occurrences call for additional scrutiny toward the less egregious features of the data, such as bimodal distributions. In a practical setting, this would need to be addressed before moving forward.

*Data Preparation.* Pushing forward with the suspicious data, we proceeded cautiously by creating a factor that represented particular cohorts within the data. We created new variables based on our knowledge of baseball.

*Modeling*. In the first model, we incorporated all the remaining features of the data. In the second model, we trained separate models for each level of SO\_FACTOR, and combined these models into an overall piecewise linear model. In the third model, we improved on the first model by dropping one of a pair of highly correlated variables.

Below, we restate the RMSE for each model constructed.

## [1] "RMSE for model 1, missing values imputed:"

## [1] 15.56565

## [1] "RMSE for model 2, missing values imputed:"

## [1] 15.42546

## [1] "RMSE for model 3, missing values imputed:"

## [1] 15.41436

## [1] "RMSE for model 1, missing values and outliers imputed:"

## [1] 14.58961

## [1] "RMSE for model 2, missing values and outliers imputed:"

## [1] 14.82223

## [1] "RMSE for model 3, missing values and outliers imputed:"

## [1] 14.59399

The best model is Model 3 with missing values and outliers imputed. It uses the full training set in a single model, unlike Model 2. Its predictors and the intercept are highly significant, which is not true for every variable in Model 1. We believe the factors added to our model serve as an effective attempt to counteract the flaws in this data set, and the engineered features made the model more efficient and improved performance. The strategy of imputing outliers both decreased RMSE and improved our residual plots.