Data621-hw1

Joseph Elikishvili

February 26, 2018

## 1. Introduction

In this paper we will explore the set containing 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. We will build several linear regression models on the training data to predict number of wins for the team.

## 2. Data Exploration

First we will load the data and review the dimensions of the dataset

## [1] 2276 17

So we have a total of 2276 records and 17 variables.

Lets Review the data

## [1] "INDEX" "TARGET\_WINS" "TEAM\_BATTING\_H"   
## [4] "TEAM\_BATTING\_2B" "TEAM\_BATTING\_3B" "TEAM\_BATTING\_HR"   
## [7] "TEAM\_BATTING\_BB" "TEAM\_BATTING\_SO" "TEAM\_BASERUN\_SB"   
## [10] "TEAM\_BASERUN\_CS" "TEAM\_BATTING\_HBP" "TEAM\_PITCHING\_H"   
## [13] "TEAM\_PITCHING\_HR" "TEAM\_PITCHING\_BB" "TEAM\_PITCHING\_SO"  
## [16] "TEAM\_FIELDING\_E" "TEAM\_FIELDING\_DP"

It appears we have Index which is just the index of the data and can be removed and we have our target variable Target\_Wins and a total of 15 predictor variables we will be working with.

Next we will scan for missing values as those can affect the process

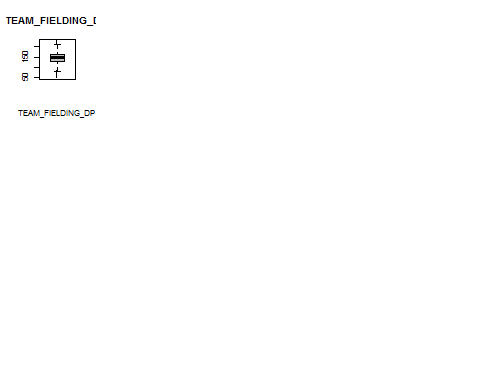
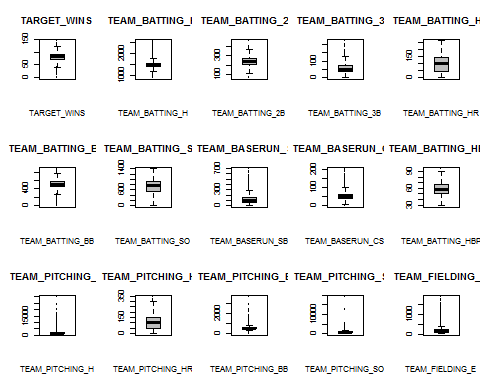
## INDEX TARGET\_WINS TEAM\_BATTING\_H TEAM\_BATTING\_2B   
## 0 0 0 0   
## TEAM\_BATTING\_3B TEAM\_BATTING\_HR TEAM\_BATTING\_BB TEAM\_BATTING\_SO   
## 0 0 0 102   
## TEAM\_BASERUN\_SB TEAM\_BASERUN\_CS TEAM\_BATTING\_HBP TEAM\_PITCHING\_H   
## 131 772 2085 0   
## TEAM\_PITCHING\_HR TEAM\_PITCHING\_BB TEAM\_PITCHING\_SO TEAM\_FIELDING\_E   
## 0 0 102 0   
## TEAM\_FIELDING\_DP   
## 286

We can see that we have a reasonable number of missing values of the most of the predictors with the exception of TEAM\_BATTING\_HBP ( 2085 ) and TEAM\_BASERUN\_CS ( 772). These are significant numbers of missing data points considering we have a total 2276 records. One option is to remove the 2 columns, but we will attempt to impute the missing data and see if it can still be useful to us before removing the columns.

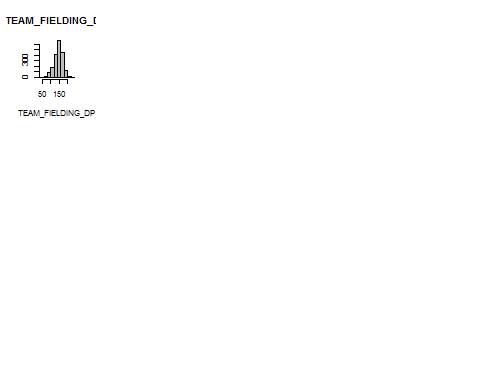
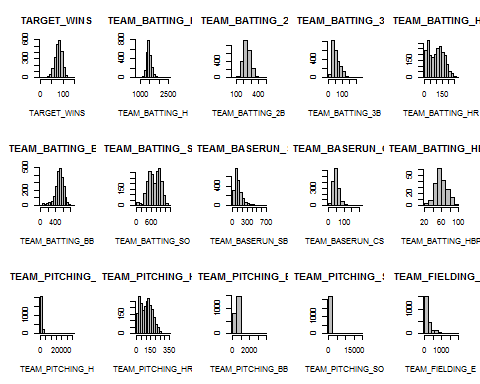
Next we will review the predictors given to us to better understand the data we are dealing with. The table gives us a basic overview of the data

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
| TARGET\_WINS | 1 | 2276 | 80.79 | 15.75 | 82.0 | 81.31 | 14.83 | 0 | 146 | 146 | -0.40 | 1.03 | 0.33 |
| TEAM\_BATTING\_H | 2 | 2276 | 1469.27 | 144.59 | 1454.0 | 1459.04 | 114.16 | 891 | 2554 | 1663 | 1.57 | 7.28 | 3.03 |
| TEAM\_BATTING\_2B | 3 | 2276 | 241.25 | 46.80 | 238.0 | 240.40 | 47.44 | 69 | 458 | 389 | 0.22 | 0.01 | 0.98 |
| TEAM\_BATTING\_3B | 4 | 2276 | 55.25 | 27.94 | 47.0 | 52.18 | 23.72 | 0 | 223 | 223 | 1.11 | 1.50 | 0.59 |
| TEAM\_BATTING\_HR | 5 | 2276 | 99.61 | 60.55 | 102.0 | 97.39 | 78.58 | 0 | 264 | 264 | 0.19 | -0.96 | 1.27 |
| TEAM\_BATTING\_BB | 6 | 2276 | 501.56 | 122.67 | 512.0 | 512.18 | 94.89 | 0 | 878 | 878 | -1.03 | 2.18 | 2.57 |
| TEAM\_BATTING\_SO | 7 | 2174 | 735.61 | 248.53 | 750.0 | 742.31 | 284.66 | 0 | 1399 | 1399 | -0.30 | -0.32 | 5.33 |
| TEAM\_BASERUN\_SB | 8 | 2145 | 124.76 | 87.79 | 101.0 | 110.81 | 60.79 | 0 | 697 | 697 | 1.97 | 5.49 | 1.90 |
| TEAM\_BASERUN\_CS | 9 | 1504 | 52.80 | 22.96 | 49.0 | 50.36 | 17.79 | 0 | 201 | 201 | 1.98 | 7.62 | 0.59 |
| TEAM\_BATTING\_HBP | 10 | 191 | 59.36 | 12.97 | 58.0 | 58.86 | 11.86 | 29 | 95 | 66 | 0.32 | -0.11 | 0.94 |
| TEAM\_PITCHING\_H | 11 | 2276 | 1779.21 | 1406.84 | 1518.0 | 1555.90 | 174.95 | 1137 | 30132 | 28995 | 10.33 | 141.84 | 29.49 |
| TEAM\_PITCHING\_HR | 12 | 2276 | 105.70 | 61.30 | 107.0 | 103.16 | 74.13 | 0 | 343 | 343 | 0.29 | -0.60 | 1.28 |
| TEAM\_PITCHING\_BB | 13 | 2276 | 553.01 | 166.36 | 536.5 | 542.62 | 98.59 | 0 | 3645 | 3645 | 6.74 | 96.97 | 3.49 |
| TEAM\_PITCHING\_SO | 14 | 2174 | 817.73 | 553.09 | 813.5 | 796.93 | 257.23 | 0 | 19278 | 19278 | 22.17 | 671.19 | 11.86 |
| TEAM\_FIELDING\_E | 15 | 2276 | 246.48 | 227.77 | 159.0 | 193.44 | 62.27 | 65 | 1898 | 1833 | 2.99 | 10.97 | 4.77 |
| TEAM\_FIELDING\_DP | 16 | 1990 | 146.39 | 26.23 | 149.0 | 147.58 | 23.72 | 52 | 228 | 176 | -0.39 | 0.18 | 0.59 |

Next we will create a boxplot of each of the predictors to visualize the variability of the data



And finally we will review the histograms of the predictors to see the distribution and the skew.



Next we will impute the missing data. Even though we have 2 predictors that have a lot of missing values, we will use them nevertheless and review the results. When we get to model selection, we will try a model with the predictors that have been heavily imputed and compare to the model without them.

For imputation we will use missForest package which will choose the best suited method and will impute the missing data.

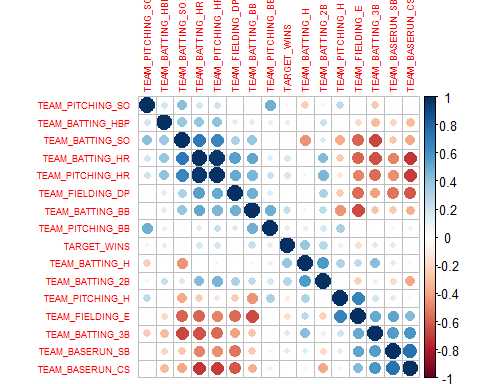
## missForest iteration 1 in progress...done!  
## missForest iteration 2 in progress...done!  
## missForest iteration 3 in progress...done!  
## missForest iteration 4 in progress...done!

Lets review the imputed data and scan for missing values.

## TARGET\_WINS TEAM\_BATTING\_H TEAM\_BATTING\_2B TEAM\_BATTING\_3B   
## 0 0 0 0   
## TEAM\_BATTING\_HR TEAM\_BATTING\_BB TEAM\_BATTING\_SO TEAM\_BASERUN\_SB   
## 0 0 0 0   
## TEAM\_BASERUN\_CS TEAM\_BATTING\_HBP TEAM\_PITCHING\_H TEAM\_PITCHING\_HR   
## 0 0 0 0   
## TEAM\_PITCHING\_BB TEAM\_PITCHING\_SO TEAM\_FIELDING\_E TEAM\_FIELDING\_DP   
## 0 0 0 0

We no longer have any missing values and we are clear to proceed further.

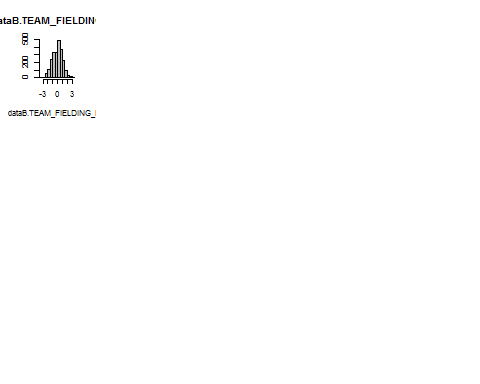
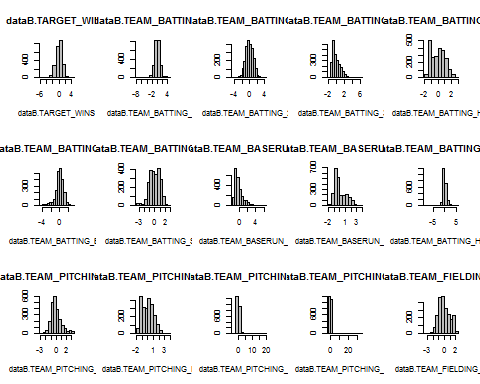
Next we will create the correlation plot to visually identify correlated predictors.



## 3. Data Preparation

Next we will use the Boxcos transformation to normalize the data. Once normalized, this will allow us to better work with the data. We will use Boxcox transformation to automatically select the appropriate transformation algorytm for our data

Next we will visualize data to see how well it was normalized after boxcox transformation and compare the results to the histograms before we applied the transformation.



We can see that boxcox transformation has had a very positive result on the dataset and data is now normalized for the most part.

## 4. Build Model

First we will create a basic model consisting of all imputed data

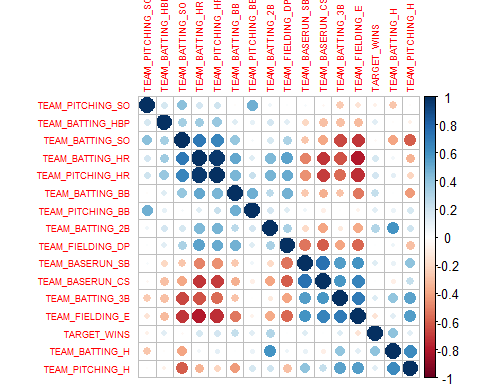
##   
## Call:  
## lm(formula = TARGET\_WINS ~ ., data = datanorm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9256 -0.5039 -0.0037 0.5220 3.4949   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.656e-11 1.672e-02 0.000 1.000000   
## TEAM\_BATTING\_H 4.204e-01 3.615e-02 11.628 < 2e-16 \*\*\*  
## TEAM\_BATTING\_2B -3.470e-02 2.697e-02 -1.287 0.198354   
## TEAM\_BATTING\_3B 1.297e-01 3.009e-02 4.310 1.70e-05 \*\*\*  
## TEAM\_BATTING\_HR -3.129e-03 1.070e-01 -0.029 0.976669   
## TEAM\_BATTING\_BB 2.344e-01 3.692e-02 6.350 2.60e-10 \*\*\*  
## TEAM\_BATTING\_SO -4.313e-01 4.217e-02 -10.228 < 2e-16 \*\*\*  
## TEAM\_BASERUN\_SB 2.666e-01 3.014e-02 8.846 < 2e-16 \*\*\*  
## TEAM\_BASERUN\_CS 1.626e-01 3.429e-02 4.743 2.24e-06 \*\*\*  
## TEAM\_BATTING\_HBP 1.260e-01 1.860e-02 6.770 1.64e-11 \*\*\*  
## TEAM\_PITCHING\_H -2.053e-01 4.040e-02 -5.081 4.06e-07 \*\*\*  
## TEAM\_PITCHING\_HR 2.794e-01 9.699e-02 2.881 0.004005 \*\*   
## TEAM\_PITCHING\_BB -1.329e-01 3.641e-02 -3.650 0.000268 \*\*\*  
## TEAM\_PITCHING\_SO 1.687e-01 3.077e-02 5.482 4.67e-08 \*\*\*  
## TEAM\_FIELDING\_E -4.772e-01 3.777e-02 -12.635 < 2e-16 \*\*\*  
## TEAM\_FIELDING\_DP -1.662e-01 2.349e-02 -7.077 1.96e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7976 on 2260 degrees of freedom  
## Multiple R-squared: 0.368, Adjusted R-squared: 0.3638   
## F-statistic: 87.73 on 15 and 2260 DF, p-value: < 2.2e-16

Next we will remove the heavily imputed predictors and compare the results

##   
## Call:  
## lm(formula = TARGET\_WINS ~ . - TEAM\_BASERUN\_CS - TEAM\_BATTING\_HBP,   
## data = datanorm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9452 -0.5053 0.0070 0.5355 3.5244   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.657e-11 1.695e-02 0.000 1.000000   
## TEAM\_BATTING\_H 4.263e-01 3.663e-02 11.637 < 2e-16 \*\*\*  
## TEAM\_BATTING\_2B -4.307e-02 2.727e-02 -1.580 0.114355   
## TEAM\_BATTING\_3B 1.385e-01 3.024e-02 4.579 4.92e-06 \*\*\*  
## TEAM\_BATTING\_HR -1.174e-01 1.070e-01 -1.097 0.272548   
## TEAM\_BATTING\_BB 2.527e-01 3.729e-02 6.778 1.55e-11 \*\*\*  
## TEAM\_BATTING\_SO -3.864e-01 4.236e-02 -9.121 < 2e-16 \*\*\*  
## TEAM\_BASERUN\_SB 3.242e-01 2.608e-02 12.430 < 2e-16 \*\*\*  
## TEAM\_PITCHING\_H -2.050e-01 4.048e-02 -5.064 4.44e-07 \*\*\*  
## TEAM\_PITCHING\_HR 3.373e-01 9.803e-02 3.441 0.000591 \*\*\*  
## TEAM\_PITCHING\_BB -1.471e-01 3.674e-02 -4.002 6.47e-05 \*\*\*  
## TEAM\_PITCHING\_SO 1.835e-01 3.105e-02 5.911 3.92e-09 \*\*\*  
## TEAM\_FIELDING\_E -4.797e-01 3.817e-02 -12.567 < 2e-16 \*\*\*  
## TEAM\_FIELDING\_DP -2.032e-01 2.320e-02 -8.758 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8086 on 2262 degrees of freedom  
## Multiple R-squared: 0.3499, Adjusted R-squared: 0.3462   
## F-statistic: 93.66 on 13 and 2262 DF, p-value: < 2.2e-16

So it appears that imputing the 2 columns did not hurt our model in fact the extra data slightly helps the model so we will leave those in.

For the next model we will try to eliminate some of the highly correlated predictors, so we will run correlation matrix again on the normalized data



We will remove TEAM\_BATTING\_HR and TEAM\_FIELDING\_E as they are very highly correlated to other predictors

##   
## Call:  
## lm(formula = TARGET\_WINS ~ . - TEAM\_BATTING\_HR - TEAM\_FIELDING\_E,   
## data = datanorm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.7813 -0.5275 0.0162 0.5582 3.3637   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.531e-11 1.729e-02 0.000 1.000   
## TEAM\_BATTING\_H 4.677e-01 3.623e-02 12.907 < 2e-16 \*\*\*  
## TEAM\_BATTING\_2B 2.995e-02 2.738e-02 1.094 0.274   
## TEAM\_BATTING\_3B 3.647e-02 2.946e-02 1.238 0.216   
## TEAM\_BATTING\_BB 2.709e-01 3.505e-02 7.727 1.64e-14 \*\*\*  
## TEAM\_BATTING\_SO -3.267e-01 4.276e-02 -7.639 3.22e-14 \*\*\*  
## TEAM\_BASERUN\_SB 2.154e-01 3.077e-02 6.999 3.37e-12 \*\*\*  
## TEAM\_BASERUN\_CS 1.364e-01 3.497e-02 3.902 9.84e-05 \*\*\*  
## TEAM\_BATTING\_HBP 1.369e-01 1.919e-02 7.136 1.29e-12 \*\*\*  
## TEAM\_PITCHING\_H -3.260e-01 3.895e-02 -8.369 < 2e-16 \*\*\*  
## TEAM\_PITCHING\_HR 3.643e-01 3.672e-02 9.921 < 2e-16 \*\*\*  
## TEAM\_PITCHING\_BB -1.431e-01 3.367e-02 -4.250 2.23e-05 \*\*\*  
## TEAM\_PITCHING\_SO 1.648e-01 3.015e-02 5.465 5.14e-08 \*\*\*  
## TEAM\_FIELDING\_DP -1.269e-01 2.406e-02 -5.274 1.46e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.825 on 2262 degrees of freedom  
## Multiple R-squared: 0.3233, Adjusted R-squared: 0.3194   
## F-statistic: 83.14 on 13 and 2262 DF, p-value: < 2.2e-16

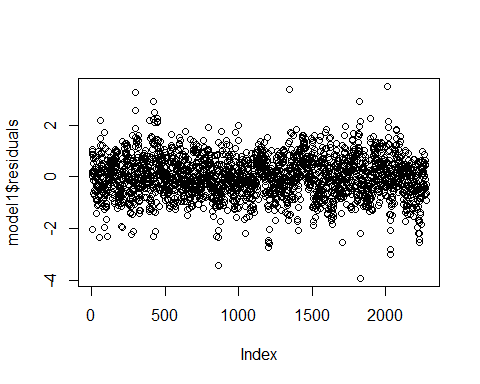
It appears that removing highly correlated predictors also does not help our model.

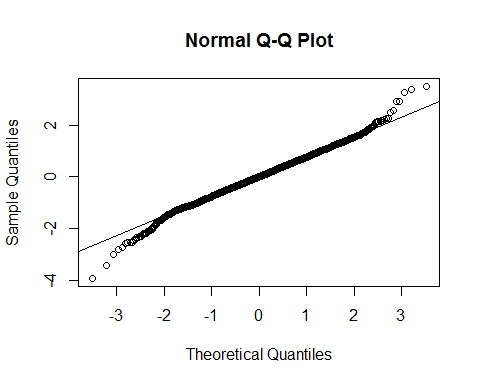
After comparing RMSE, R-squared and F-statistics of the 3 models, we will go ahead and use the model1 it is the simplest model and uses all the predictors, but since we did not find any significant benefit with other models, We will use the simplest model containing all the predictors.

## 5. Select Model

We decided to select the final model as it is the model in its purest form containing all the predictors, we tried to omit imputed data, drop some of the highly correlated predictors and removing some of the less relevant predictors, but did not see much improvement, therefore we decided to stay with the base model.

Next we will review residual plot and qqplot of the model. We want to make sure the residuals are random and the variance is constant.





## missForest iteration 1 in progress...done!  
## missForest iteration 2 in progress...done!  
## missForest iteration 3 in progress...done!

## References

Complete code is located at: <https://github.com/jelikish/Cuny1/blob/master/Spring2018/621/hw1>