Homework 1

Group 1

Contents

Prepared for:

Dr. Nathan Bastian
City University of New York - Data 621

Prepared by:

Group 1

Senthil Dhanapal
Yadu Chittampalli
Christophe Hunt

1 Introduction

The ability to analyze and predict performance of a professional baseball team using many dimensions is critical to competitive success for our organization. Therefore, we have analyzed the records of numerous professional baseball team from the years 1871 to 2006. Our hope is that the following report and the resulting predictive models will better inform the organization and assist in making data driven decisions moving forward.

"The goal of a baseball team is to win more games than any other team. Since one team has very little control over the number of games other teams win, the goal is essentially to win as many games as possible. Therefore, it is of interest to measure the player's contribution to the team's wins." Grabiner, B. D. [1] While we do not have the variables at the player's individual contribution level, we do have the entire teams contributions as an aggregate and will analyze that information.

2 Statement of the Problem

The purpose of this report is to determine the batting, baserun, pitching, and fielding effects on a baseball team's ability to win.

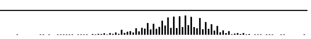
3 Data Exploration

Note that 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. The following Table 1 - Descriptive Statistics provides the detailed descriptive statistics regarding our variable of interest - Number of Wins and our possible explanatory variables.

We noted that several variables were missing a nontrivial amount of observations and these variables are Strikeouts by batters, Stolen Bases, Caught stealing, Batters hit by pitch (get a free base), Strikeouts by pitcher, and Double plays. So we will need to address the missing values for further analysis.

Table 1 : Descriptive Statistics
16 Variables 2276 Observations

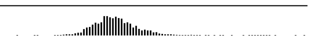
Number of wins



n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
2276	0	108	81	82	15.8	54	104

`lowest :   0  12  14  17  21, highest: 128 129 134 135 146`

Base Hits by batters (1B,2B,3B,HR)



n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
2276	0	569	1469	1454	144.6	1282	1695

`lowest :  891  992 1009 1116 1122, highest: 2333 2343 2372 2496 2554`

Doubles by batters (2B)

n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
2276	0	240	241	238	46.8	167	320

`lowest :  69 112 113 118 123, highest: 382 392 393 403 458`

[1](Grabiner, B. D. (n.d.). The Sabermetric Manifesto. Retrieved September 10, 2016 from http://seanlahman.com/baseball-archive/sabermetrics/sabermetric-manifesto/)

Triples by batters (3B)

n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
2276	0	144	55	47	27.9	23	108

```
lowest :   0   8   9  11  12, highest: 166 190 197 200 223
```

Homeruns by batters (4B)

n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
2276	0	243	100	102	60.5	14	199

```
lowest :   0   3   4   5   6, highest: 247 249 257 260 264
```

Walks by batters

n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
2276	0	533	502	512	122.7	248	670

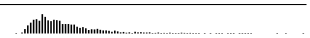
```
lowest :   0  12  29  34  45, highest: 815 819 824 860 878
```

Strikeouts by batters

n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
2174	102	822	736	750	248.5	359	1103

```
lowest :    0   66   67   72   74, highest: 1303 1320 1326 1335 1399
```

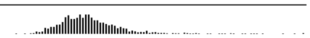
Stolen bases



n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
2145	131	348	125	101	87.8	35	302

```
lowest :   0  14  18  19  20, highest: 562 567 632 654 697
```

Caught stealing



n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
1504	772	128	53	49	23	24	91

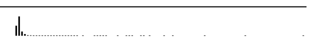
```
lowest :   0   7  11  12  14, highest: 171 186 193 200 201
```

Batters hit by pitch (get a free base)

n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
191	2085	55	59	58	13	40	82

```
lowest : 29 30 35 38 39, highest: 87 88 89 90 95
```

Hits allowed



n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
2276	0	843	1779	1518	1406.8	1316	2563

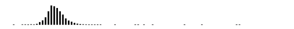
```
lowest :  1137  1168  1184  1187  1202
highest: 16038 16871 20088 24057 30132
```

Homeruns allowed

n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
2276	0	256	106	107	61.3	18	209

```
lowest :   0   3   4   5   6, highest: 291 297 301 320 343
```

Walks allowed



n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
2276	0	535	553	536.5	166.4	377	757

```
lowest :    0  119  124  131  140, highest: 2169 2396 2840 2876 3645
```

Strikeouts by pitchers

n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
2174	102	823	818	813.5	553.1	421	1173

```
lowest :     0   181   205   208   252
highest:  3450  4224  5456 12758 19278
```

Errors

n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
2276	0	549	246	159	227.8	100	716

```
lowest :   65   66   68   72   74, highest: 1567 1728 1740 1890 1898
```

Double Plays

n	missing	unique	Mean	Median	SD	.05 freq	.95 freq
1990	286	144	146	149	26.2	98	186

```
lowest :  52  64  68  71  72, highest: 215 218 219 225 228
```

3.1 Imputing Missing Values

In order to address the missing values in our variables we used a nonparametric imputation method (Random Forest) to impute missing values. We chose a nonparametric method due to several variables having significant skew and greater than expected kurtosis values.

3.2 Correlation Matrix

After competing the imputation, we can implement a correlation matrix to better understand the correlation between variables in the data set. The below matrix is the results and interestingly, Number of Wins appears to be most correlated to Base Hits by batters (1B,2B,3B,HR).

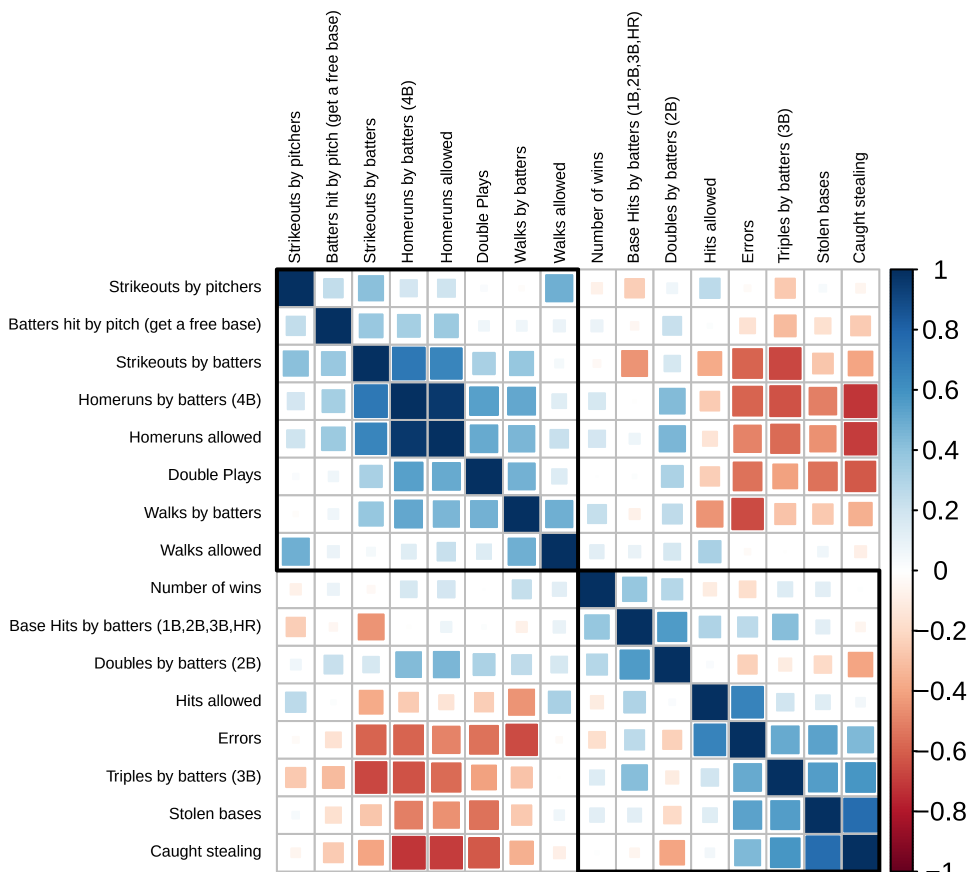


Figure 1: Correlation Plot of Training Data Set with imputed values

4 Data Preparation

First, we chose to eliminate two variables that had a significant number of missing data points. These variables were Batters hit by pitch (get a free base) and Caught stealing, which were missing 91.6% and 33.9% respectively.

Missing values in the remaining columns had been imputed using the random forest method as previous discussed in section 3.1.

The rows containing all of the outliers were removed from the original dataset to reduce the skewness. After all of these rows were removed, the skew in the number of walks allowed was significantly reduced to below 1.

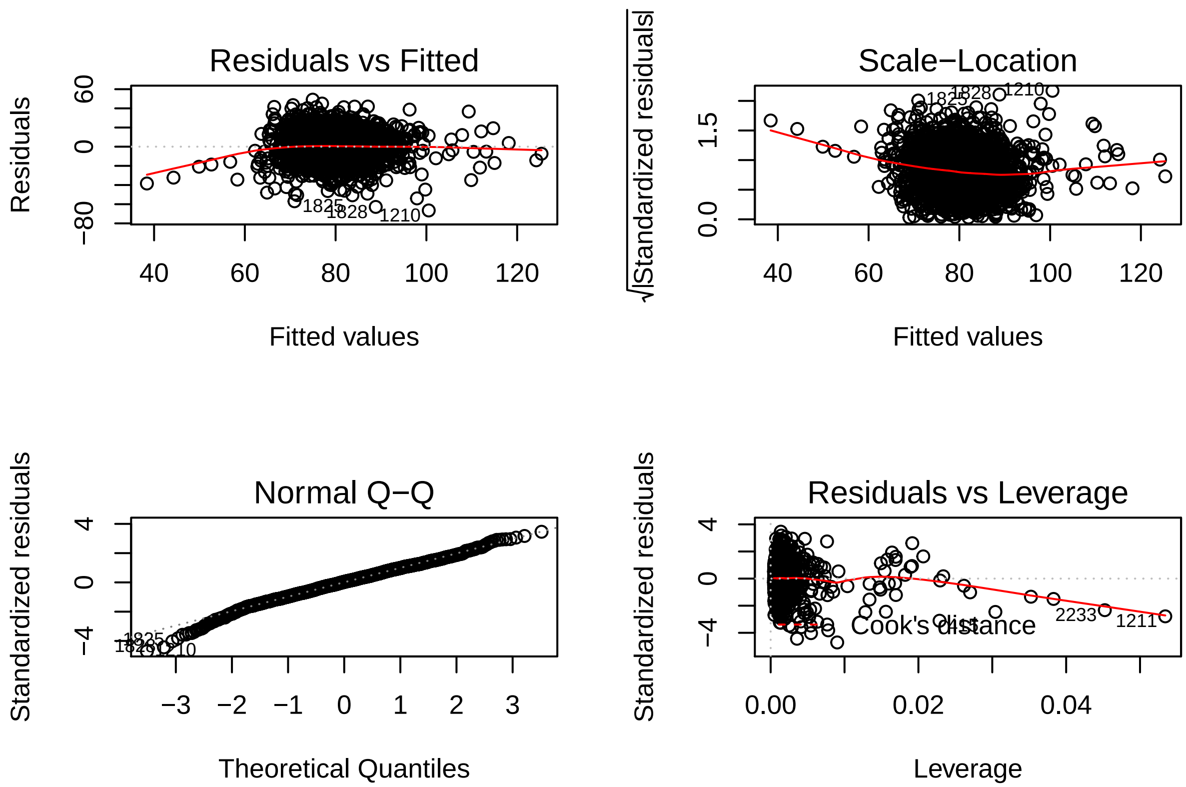
The Box-Cox transformation was done on the number of base hits by batters. As a result of this transformation, the skew in the number of base hits by batters was significantly reduced to below 1.

5 Models Built

```
lmfit <- lm(data = imputed_df,
            `Number of wins` ~ `Base Hits by batters (1B,2B,3B,HR)` + `Homeruns by batters (4B)` +  sqrt
summary(lmfit)
```

```
##
## Call:
## lm(formula = `Number of wins` ~ `Base Hits by batters (1B,2B,3B,HR)` +
##     `Homeruns by batters (4B)` + sqrt(`Strikeouts by batters`),
##     data = imputed_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.519  -9.312   0.471   9.544  49.048
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -5.869802   5.424862  -1.082  0.27936
## `Base Hits by batters (1B,2B,3B,HR)`  0.049707   0.002634  18.870  < 2e-16
## `Homeruns by batters (4B)`            0.021622   0.007497   2.884  0.00396
## sqrt(`Strikeouts by batters`)         0.432401   0.098461   4.392 1.18e-05
##
## (Intercept)
## `Base Hits by batters (1B,2B,3B,HR)` ***
## `Homeruns by batters (4B)`           **
## sqrt(`Strikeouts by batters`)        ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.19 on 2272 degrees of freedom
## Multiple R-squared:   0.19,  Adjusted R-squared:  0.1889
## F-statistic: 177.6 on 3 and 2272 DF,  p-value: < 2.2e-16
```

```r
layout(matrix(c(1,2,3,4),2,2))
plot(lmfit)
```



Using the training data set, build at least three different multiple linear regression models, using different variables (or the same variables with different transformations). Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.). Since you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done. Discuss the coefficients in the models, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

6 Selected Model

Decide on the criteria for selecting the best multiple linear regression model. Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model. For the multiple linear regression model, will you use a metric such as Adjusted R2 , RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R2 , (c) F-statistic, and (d) residual plots. Make predictions using the evaluation data set.

7 Appendix A

7.1 Session Info

- R version 3.3.1 (2016-06-21), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=English_United States.1252`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, utils
- Other packages: bibtex 0.4.0, corrplot 0.77, doParallel 1.0.10, dplyr 0.5.0, e1071 1.6-7, foreach 1.4.3, Formula 1.2-1, ggplot2 2.1.0, Hmisc 3.17-4, iterators 1.0.8, itertools 0.1-3, knitcitations 1.0.7, knitr 1.14, lattice 0.20-34, magrittr 1.5, missForest 1.4, pacman 0.4.1, plyr 1.8.4, randomForest 4.6-12, rJava 0.9-8, scales 0.4.0, stringr 1.1.0, survival 2.39-5, tidyr 0.6.0, xlsx 0.5.7, xlsxjars 0.6.1
- Loaded via a namespace (and not attached): acepack 1.3-3.3, assertthat 0.1, bitops 1.0-6, chron 2.3-47, class 7.3-14, cluster 2.0.4, codetools 0.2-14, colorspace 1.2-6, data.table 1.9.6, DBI 0.5-1, digest 0.6.10, evaluate 0.9, foreign 0.8-67, formatR 1.4, grid 3.3.1, gridExtra 2.2.1, gtable 0.2.0, htmltools 0.3.5, httr 1.2.1, latticeExtra 0.6-28, lazyeval 0.2.0, lubridate 1.6.0, Matrix 1.2-7.1, munsell 0.4.3, nnet 7.3-12, R6 2.1.3, RColorBrewer 1.1-2, Rcpp 0.12.7, RCurl 1.95-4.8, RefManageR 0.11.0, RJSONIO 1.3-0, rmarkdown 1.0, rpart 4.1-10, splines 3.3.1, stringi 1.1.1, tibble 1.2, tools 3.3.1, XML 3.98-1.4, yaml 2.1.13

7.2 Citations

7.3 Data Dictionary

VARIABLE.NAME..	DEFINITION	THEORETICAL.EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	NA
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

VARIABLE.NAME..	DEFINITION	THEORETICAL.EFFECT
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

7.4 R source code