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 variables at the individual player's 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

Please 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 provides the detailed descriptive statistics regarding our variable of interest - Number of Wins and our possible explanatory variables.

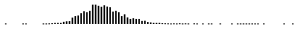
Table 1 : Descriptive Statistics
16 Variables 2276 Observations

Number of wins

n	missing	unique	Mean	median	sd	.05 freq	.95 freq	skewness	kurtosis
2276	0	108	81	82	15.8	54	104	-0.4	1

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	skewness	kurtosis
2276	0	569	1469	1454	144.6	1282	1695	1.6	7.3

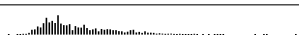
lowest : 891 992 1009 1116 1122, highest: 2333 2343 2372 2496 2554

Doubles by batters (2B)

n	missing	unique	Mean	median	sd	.05 freq	.95 freq	skewness	kurtosis
2276	0	240	241	238	46.8	167	320	0.2	0

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

Triples by batters (3B)



n	missing	unique	Mean	median	sd	.05 freq	.95 freq	skewness	kurtosis
2276	0	144	55	47	27.9	23	108	1.1	1.5

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

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

Homeruns by batters (4B)

n	missing	unique	Mean	median	sd	.05 freq	.95 freq	skewness	kurtosis
2276	0	243	100	102	60.5	14	199	0.2	-1

```
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	skewness	kurtosis
2276	0	533	502	512	122.7	248	670	-1	2.2

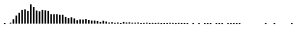
```
lowest :   0  12  29  34  45, highest: 815 819 824 860 878
```

Strikeouts by batters

n	missing	unique	Mean	median	sd	.05 freq	.95 freq	skewness	kurtosis
2174	102	822	736	750	248.5	359	1103	-0.3	-0.3

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

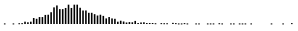
Stolen bases



n	missing	unique	Mean	median	sd	.05 freq	.95 freq	skewness	kurtosis
2145	131	348	125	101	87.8	35	302	2	5.5

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

Caught stealing



n	missing	unique	Mean	median	sd	.05 freq	.95 freq	skewness	kurtosis
1504	772	128	53	49	23	24	91	2	7.6

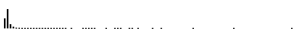
```
lowest :   0   7  11  12  14, highest: 171 186 193 200 201
```

Batters hit by pitch (get a free base)

n	missing	unique	Mean	median	sd	.05 freq	.95 freq	skewness	kurtosis
191	2085	55	59	58	13	40	82	0.3	-0.1

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

Hits allowed



n	missing	unique	Mean	median	sd	.05 freq	.95 freq	skewness	kurtosis
2276	0	843	1779	1518	1406.8	1316	2563	10.3	141.8

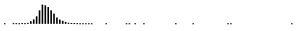
```
lowest :  1137  1168  1184  1187  1202
highest: 16038 16871 20088 24057 30132
```

Homeruns allowed

n	missing	unique	Mean	median	sd	.05 freq	.95 freq	skewness	kurtosis
2276	0	256	106	107	61.3	18	209	0.3	-0.6

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

Walks allowed



n	missing	unique	Mean	median	sd	.05 freq	.95 freq	skewness	kurtosis
2276	0	535	553	536.5	166.4	377	757	6.7	97

```
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	skewness	kurtosis
2174	102	823	818	813.5	553.1	421	1173	22.2	671.2

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

Errors

n	missing	unique	Mean	median	sd	.05 freq	.95 freq	skewness	kurtosis
2276	0	549	246	159	227.8	100	716	3	11

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

Double Plays

n	missing	unique	Mean	median	sd	.05 freq	.95 freq	skewness	kurtosis
1990	286	144	146	149	26.2	98	186	-0.4	0.2

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

Correlation Matrix

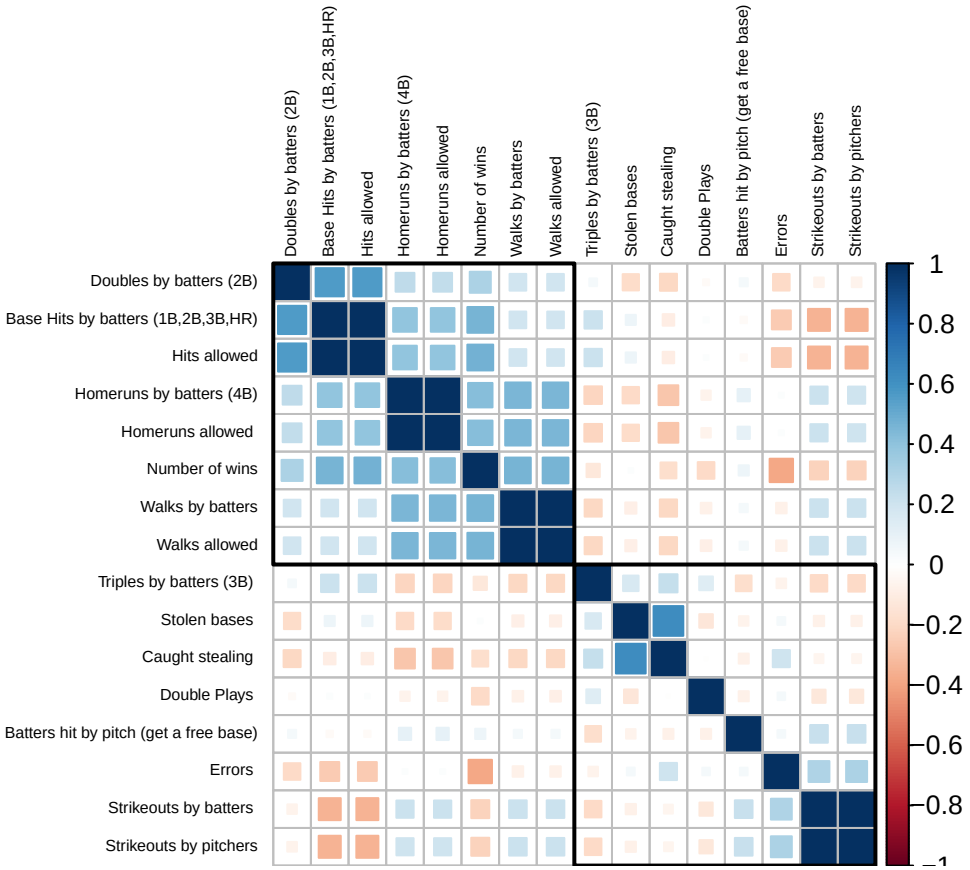


Figure 1: Correlation Plot of Training Data Set

4 Data Preparation

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations. a. Fix missing values (maybe with a Mean or Median value) b. Create flags to suggest if a variable was missing c. Transform data by putting it into buckets d. Mathematical transforms such as log or square root (or use Box-Cox) e. Combine variables (such as ratios or adding or multiplying) to create new variables

5 Models Built

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 Citations

7.2 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
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

7.3 R source code