Homework 1

Group 1

Contents

Prepared for:

Dr. Nathan Bastian

City University of New York, School of Professional Studies - 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.

Histograms of all of the variables have been plotted below so that the distribution of the data can be visualized. In the distribution for the number of walks allowed, only two bars exist due to the excessive number of outliers.

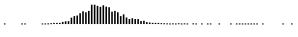
Table 1 : Descriptive Statistics
16 Variables 2276 Observations

Number of wins

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	108	81	82	15.8	-0.4	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	Skew	.05 freq	.95 freq
2276	0	569	1469	1454	144.6	1.6	1282	1695

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

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

Doubles by batters (2B)

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	240	241	238	46.8	0.2	167	320

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

Triples by batters (3B)

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	144	55	47	27.9	1.1	23	108

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

Homeruns by batters (4B)

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	243	100	102	60.5	0.2	14	199

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

Walks by batters

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	533	502	512	122.7	-1	248	670

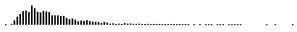
```
lowest :   0  12  29  34  45, highest: 815 819 824 860 878
```

Strikeouts by batters

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2174	102	822	736	750	248.5	-0.3	359	1103

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

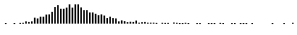
Stolen bases



n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2145	131	348	125	101	87.8	2	35	302

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

Caught stealing



n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
1504	772	128	53	49	23	2	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	Skew	.05 freq	.95 freq
191	2085	55	59	58	13	0.3	40	82

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

Hits allowed

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	843	1779	1518	1406.8	10.3	1316	2563

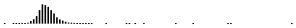
```
lowest :  1137  1168  1184  1187  1202
highest: 16038 16871 20088 24057 30132
```

Homeruns allowed

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	256	106	107	61.3	0.3	18	209

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

Walks allowed



n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	535	553	536.5	166.4	6.7	377	757

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

Strikeouts by pitchers

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2174	102	823	818	813.5	553.1	22.2	421	1173

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

Errors

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	549	246	159	227.8	3	100	716

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

Double Plays

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
1990	286	144	146	149	26.2	-0.4	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. Several variables have a significant amount of skew, which include the number of base hits by batters and the number of walks allowed. Correspondingly, these two variables had a skew of 1.57 and 6.74 respectively. Therefore, we chose a nonparametric method due to several variables having significant skew and having a non-normal distribution.

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 as expected, Number of Wins appears to be most correlated to Base Hits by batters (1B,2B,3B,HR).

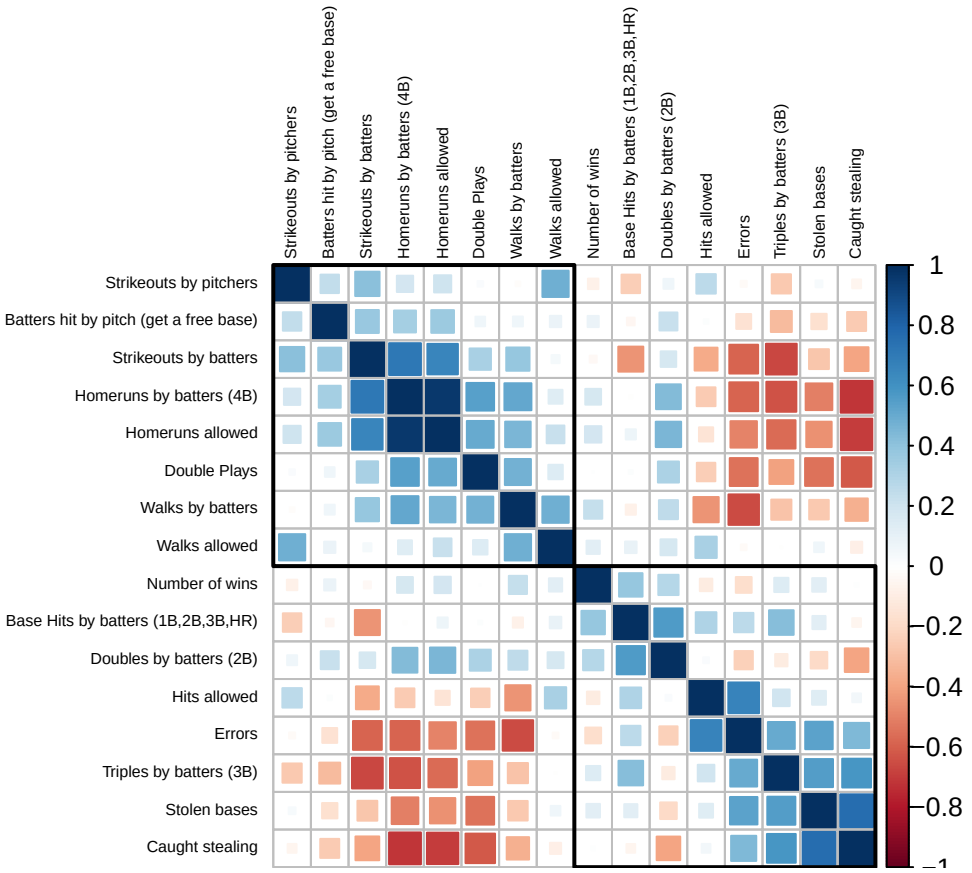


Figure 1: Correlation Plot of Training Data Set with imputed values

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.

Additionally, we reduced the data set to the following variables for modeling simplicity. Base Hits by batters (1B,2B,3B,HR), Strikeouts by batters, Walks by batters, Double plays, Walks allowed, Triples by batters (3B), and Hits allowed.

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

4.1 Outliers

We chose winsorizing as the method to address outliers. Instead of trimming values, winsorizing uses the interquantile range to replace values that are above or below the interquantile range multiplied by a factor. Those values above or below the range multiplied by the factor are then replaced with max and min value of the interquantile range. Using the factor 2.2 for winsorizing outliers is a method developed my Hoaglin and Iglewicz and published Journal of American Statistical Association in 1987[2].

The below table is the results of the winsorizing of the data.

Table 2 : Descriptive Statistics after Winsorizing Outliers
14 Variables 2276 Observations

Number of wins

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	108	81	82	15.8	-0.4	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	Skew	.05 freq	.95 freq
2276	0	533	1466	1454	144.6	1.6	1282	1695

lowest : 1116 1122 1137 1141 1142, highest: 1852 1860 1861 1864 1876

Doubles by batters (2B)

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	239	241	238	46.8	0.2	167	320

lowest : 69 112 113 118 123, highest: 378 382 392 393 403

Triples by batters (3B)

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	134	55	47	27.9	1.1	23	108

lowest : 0 8 9 11 12, highest: 143 144 145 147 151

[2]Hoaglin, D. C., and Iglewicz, B. (1987), Fine tuning some resistant rules for outlier labeling, Journal of American Statistical Association, 82, 1147-1149.

Homeruns by batters (4B)

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	243	100	102	60.5	0.2	14	199

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

Walks by batters

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	485	503	512	122.7	-1	248	670

```
lowest : 170 171 172 173 174, highest: 806 815 819 824 860
```

Strikeouts by batters

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	923	731	750	248.5	-0.3	364	1099

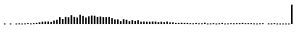
```
lowest :    0   66   67   72   74, highest: 1303 1320 1326 1335 1399
```

Stolen bases

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	440	132	101	87.8	2	36	323

```
lowest :   0  14  18  19  20, highest: 383 385 386 388 392
```

Hits allowed



n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	692	1597	49	23	2	1316	2260

```
lowest : 1137 1168 1184 1187 1202, highest: 2249 2250 2258 2259 2260
```

Homeruns allowed

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	256	106	58	13	0.3	18	209

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

Walks allowed

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	497	547	1518	1406.8	10.3	377	757

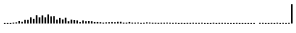
```
lowest : 188 198 203 209 218, highest: 886 890 892 896 899
```

Strikeouts by pitchers

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	913	796	107	61.3	0.3	425	1169

```
lowest :    0  181  205  208  252, highest: 1552 1561 1590 1600 1659
```

Errors



n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	352	216	536.5	166.4	6.7	100	516

```
lowest :  65  66  68  72  74, highest: 511 513 514 515 516
```

Double Plays 

n	missing	unique	Mean	Median	SD	Skew	.05 freq	.95 freq
2276	0	415	142	813.5	553.1	22.2	97	184

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

4.2 Box Cox Transformation

We choose the Box Cox transformation for the following variables to improve linearity in our model. The lambda for the Box Cox transformation of our response variable is 1.8672 which indicates that we should square the response variable to improve linearity. Also, we discovered that it was not necessary to transform Base Hits by Batters because the significance levels of the variable before and after transformation were the same.

5 Models Built

All models included Base Hits by batters (1B,2B,3B,HR) which is the most correlated variable to Number of Wins as indicated in the correlation matrix. This is expected as Base Hits are necessary to win any game.

5.1 Model 1

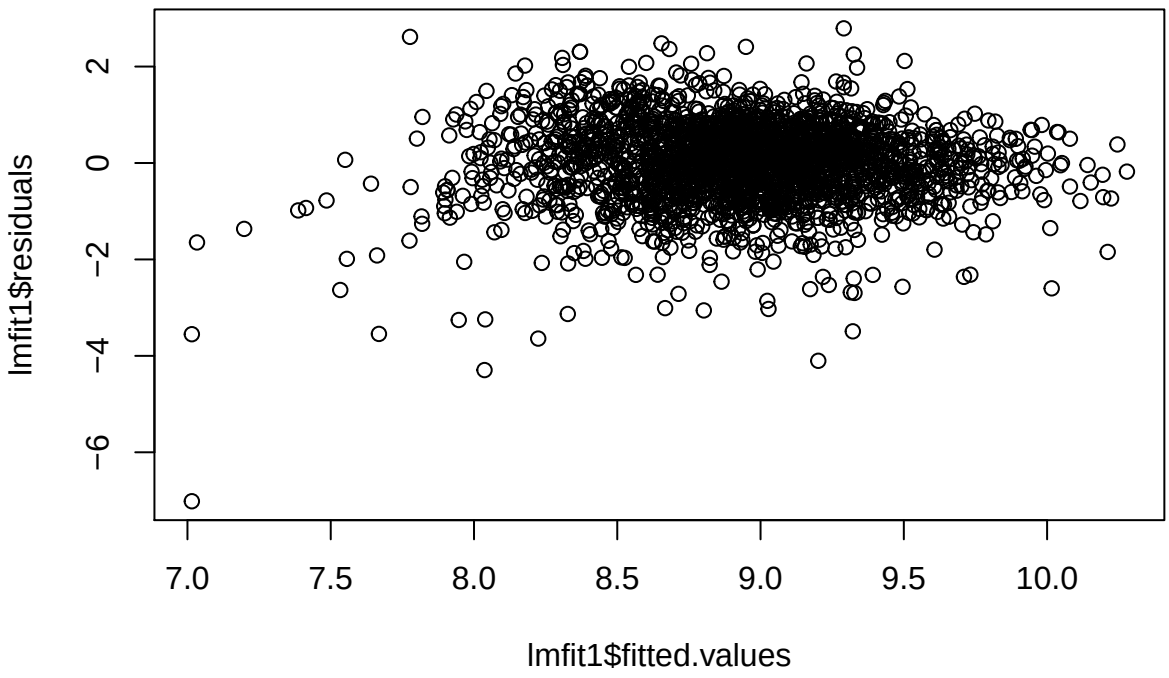
We added Walks by batters because a batter being walked would put a runner on a base and therefore in a better position to score. Additionally, Strikeouts by batters would be negatively correlated to the Number of Wins because if a batter strikes out they are not able to provide runs which are criticaly to winning.

Table 1:

	Model 1
	$\sqrt{\text{Number of wins}}$
Base Hits by batters (1B,2B,3B,HR)	0.003***
	(0.0002)
Walks by batters	0.002***
	(0.0002)
Strikeouts by batters	0.0004***
	(0.0001)
Constant	3.373***
	(0.249)
Observations	2,276
R^2	0.210
Adjusted R^2	0.209
Residual Std. Error	0.828 (df = 2272)
F Statistic	201.646*** (df = 3; 2272)
Note:	*p<0.1; **p<0.05; ***p<0.01

The F-statistic is 201.6, and the p-value indicates that this model is significant. Additionally, we see the adjusted R-squared is .2092 but unexpectedly Strikeouts by batters has a positive coefficient and all three predictors are significant.

The below plot of our fitted values against our residuals indicate that there is Heteroskadistity and showing uneven variation.



5.2 Model 2

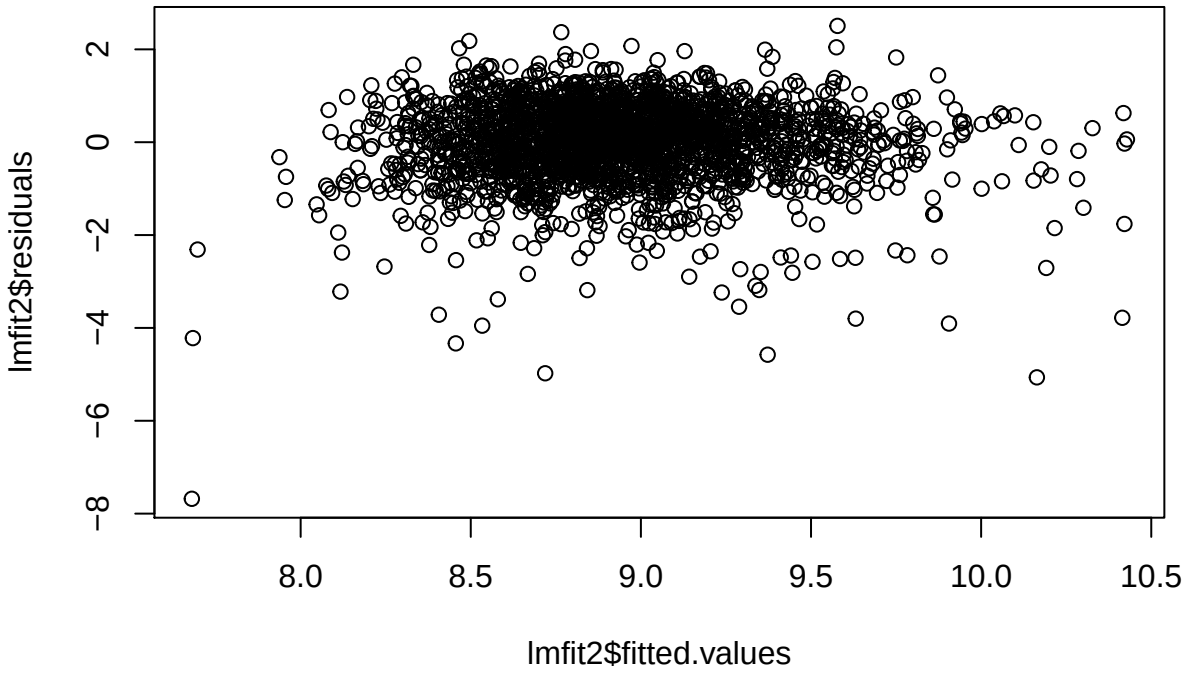
We added Walks allowed and Double Plays. The reason being that Walks allowed is possibly an indicator of poor pitching and Double Plays is an indicator of a competent infield team that prevents other teams from scoring.

The F-statistic is 141.2, and the p-value indicates that this model is significant. We see the adjusted R-squared is 0.1561, however, it was unexpected that double plays has a negative coeffiecient but it was not a significant predictor.

The below plot of our fitted values against our residuals indicate that there is Heteroskadistity and showing uneven variation.

Table 2:

	Model 2
	$\sqrt{\text{Number of wins}}$
Base Hits by batters (1B,2B,3B,HR)	0.003***
	(0.0001)
Walks allowed	0.001***
	(0.0002)
Double Plays	−0.001
	(0.001)
Constant	4.730***
	(0.224)
Observations	2,276
R^2	0.157
Adjusted R^2	0.156
Residual Std. Error	0.855 (df = 2272)
F Statistic	141.228*** (df = 3; 2272)
Note:	*p<0.1; **p<0.05; ***p<0.01



5.3 Model 3

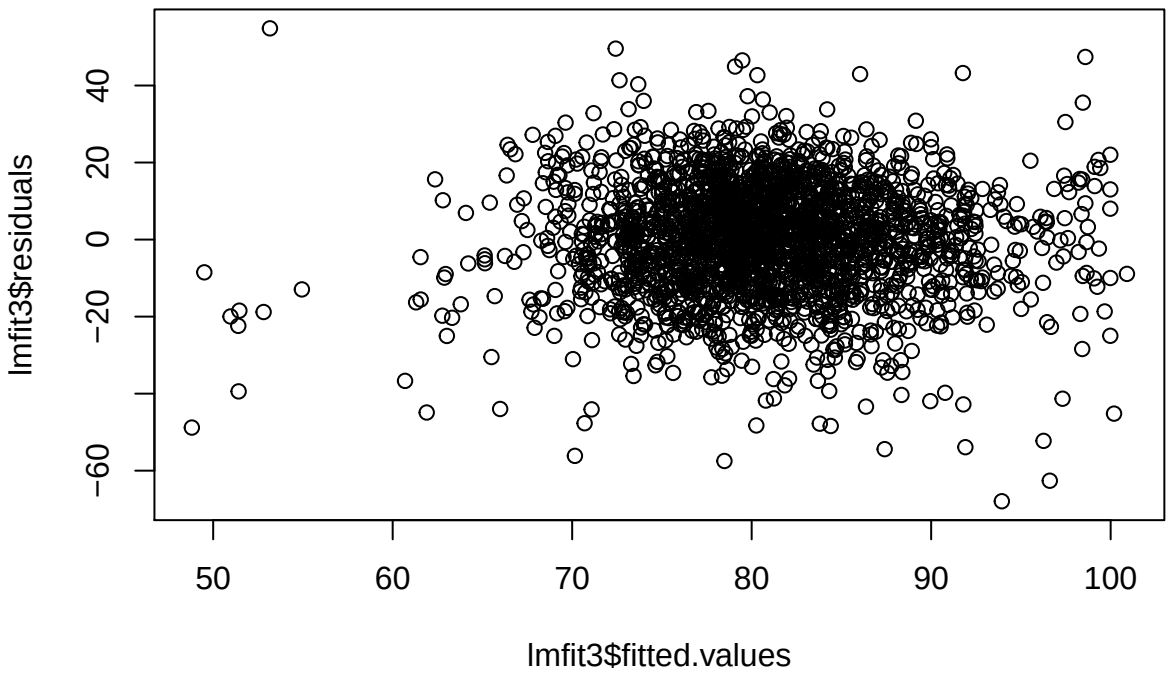
We included Triples by batters and Hits allowed. Additonally, as expected, the Hits allowed has a negative relationship to wins because hits allowed indicates the other team getting a hit and possibly scoring a point.

Table 3:

	Model 3
	$\sqrt{\text{Number of wins}}$
Base Hits by batters (1B,2B,3B,HR)	0.060^{***}
	(0.003)
Triples by batters (3B)	0.035^{***}
	(0.013)
Hits allowed	-0.013^{***}
	(0.002)
Constant	11.748^{***}
	(3.522)
Observations	2,276
R^2	0.172
Adjusted R^2	0.171
Residual Std. Error	14.341 (df = 2272)
F Statistic	157.533^{***} (df = 3; 2272)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

The F-statistic is 157.5 and based on our p-values this model is significant. The Adjusted R-squared is 0.1711. However, the Triples by batters (3B) is less significant due to the colinearity with Base Hits by batters (1B, 2B,3B, HR).

The below plot of our fitted values against our residuals indicate that there is Heteroskadistity and showing uneven variation.


lmfit3$residuals
40
20
0
−20
−60
50
60
70
80
90
100
lmfit3$fitted.values

6 Selected Model

Our residuals in each model indcated Heteroskadistity and showed uneven variation. Therefore, we chose the model with the highest Adjusted R-squared which was Model 1. The Adjusted R-squared of each model is provided in the below table.

Model	1	2	3
Adjusted R Squared	0.209	0.156	0.171
F-statistic	201.646	141.228	157.533
P Value for F-Statistic	Significant	Significant	Significant
Residual vs Fitted Constant Variation	Not good	Not good	Not good
Residual vs Fitted Curve	shows curve	shows curve	shows curve
Residual vs Fitted Curve Heteroscedasticity	yes	yes	yes

Therefore, our final model with the greatest Adjusted R-squared is:

$$\begin{aligned}\sqrt{\text{Number of wins}} = {} & 0.002994 * \text{Base Hits by batters (1B,2B,3B,HR)} \\ & + 0.001773 * \text{Walks by batters} \\ & + 0.0003899 * \text{Strikeouts by batters} \\ & + 3.372983\end{aligned}$$

Additionally, we are 95% confident that our regression values lie between the two values from the output in the below table.

	2.5 %	97.5 %
(Intercept)	2.8837578	3.8622073
`Base Hits by batters (1B,2B,3B,HR)`	0.0026985	0.0032888
`Walks by batters`	0.0014558	0.0020902
`Strikeouts by batters`	0.0002217	0.0005582

7 Prediction on Evaluation Data

Using our best performing model we used the predict R function and excluded evaluation data with missing values. The below table is our prediction results for the evaluation data set.

Identification Variable (do not use)	predictions
9	67
10	68
14	76
47	84
60	64
63	66
74	78
83	70
98	67
120	73
123	76
135	79

Identification Variable (do not use)	predictions
138	77
140	79
151	75
153	76
171	72
184	82
213	92
217	84
226	90
230	83
241	75
291	81
294	84
300	64
348	74
350	87
357	78
367	93
368	87
372	89
382	93
388	81
396	82
398	76
403	87
407	88
410	87
412	84
414	89
436	46
440	96
476	85
479	90
481	94
501	69
503	68
506	73
519	77
522	87
550	77
554	70
566	77
578	75
596	84
599	66
614	82
644	83
692	85
699	82
700	78
716	85
721	74

Identification Variable (do not use)	predictions
722	80
731	82
746	84
763	75
774	80
776	82
788	83
789	86
792	81
811	78
861	88
862	89
863	98
871	83
879	85
887	79
892	75
904	81
909	83
925	89
940	76
951	92
976	71
989	84
995	101
1000	90
1001	90
1007	83
1016	74
1027	85
1033	80
1070	82
1081	72
1084	61
1098	81
1150	84
1160	67
1169	80
1172	79
1174	89
1176	85
1178	78
1184	81
1193	89
1196	79
1199	78
1207	70
1218	83
1229	66
1241	82
1244	89
1246	71

Identification Variable (do not use)	predictions
1248	87
1249	94
1253	90
1261	81
1305	75
1314	83
1323	82
1328	70
1353	74
1363	76
1371	81
1372	80
1389	63
1421	93
1431	80
1437	76
1442	75
1450	79
1463	83
1464	84
1470	79
1471	80
1484	81
1495	64
1507	71
1514	77
1526	71
1549	83
1552	69
1556	85
1585	102
1586	105
1590	90
1591	104
1592	96
1603	88
1612	83
1634	79
1645	72
1647	81
1674	82
1687	80
1688	91
1700	83
1708	78
1713	80
1717	76
1721	76
1730	81
1737	76
1748	83
1749	84

Identification Variable (do not use)	predictions
1763	84
1768	98
1778	85
1780	92
1782	68
1784	65
1794	113
1819	72
1832	78
1833	84
1844	71
1847	76
1854	80
1855	78
1857	86
1864	78
1865	80
1869	74
1880	84
1881	78
1882	79
1894	83
1896	78
1916	79
1918	74
1921	97
1926	88
1938	77
1979	69
1982	72
1987	86
1997	82
2004	84
2011	75
2015	76
2022	81
2025	74
2027	86
2031	78
2036	102
2066	73
2073	79
2087	83
2092	82
2125	78
2162	88
2191	82
2203	83
2218	78
2221	73
2225	84
2232	74

Identification Variable (do not use)	predictions
2267	86
2291	69
2299	87
2317	86
2318	85
2353	82
2403	66
2411	87
2415	77
2424	81
2441	72
2464	82
2465	82
2472	64
2481	92
2487	43
2500	69
2501	79
2520	75
2521	79
2525	77

8 Appendix A

8.1 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, forecast 7.2, 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, stargazer 5.2, stringr 1.1.0, survival 2.39-5, tidyr 0.6.0, timeDate 3012.100, xlsx 0.5.7, xlsxjars 0.6.1, xtable 1.8-2, zoo 1.7-13
- 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, fracdiff 1.4-2, grid 3.3.1, gridExtra 2.2.1, gtable 0.2.0, highr 0.6, 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, quadprog 1.5-5, 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, tseries 0.10-35, XML 3.98-1.4, yaml 2.1.13

8.2 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

8.3 R source code

Please see Homework 1.rmd on GitHub for source code.

https://github.com/ChristopheHunt/DATA-621-Group-1/blob/master/Homework%201.Rmd.