GROUP#2 HOMEWORK# 1 - MONEY BALL

Abstract

What causes teams to have success in America's game? Using data from baseball dating back to 1871, statistical analysis is performed to make inferences on what variables make a team a winner. A model is then built to predict the number of wins in a season based on those variables.

Moneyball Project

Group2

2/12/2018

Introduction

With the aim of producing a predictive model for baseball wins in a season, we are proposing to explore, analyze and model a data set containing approximately 2000 records.

Each record represents statistics for a professional baseball team from the years 1871-2006. Each record has the performance of the team for the given year, with all the statistics adjusted to match the performance of a 162-game season.

The predictor variables considered in the linear modeling exercise to predict wins are:

Variable Name	Definition
BATTING_H	Base Hits by batters (1B,2B,3B,HR)
BATTING_2B	Doubles by batters (2B)
BATTING_3B	Triples by batters (3B)
BATTING_HR	Homeruns by batters (4B)
BATTING_BB	Walks by batters
BATTING_HBP	Batters hit by pitch
BATTING_SO	Batters hit by pitch
BASERUN_SB	Stolen bases
BASERUN_CS	Caught stealing
TFIELDING_E	Errors
FIELDING_DP	Double Plays
PITCHING_BB	Walks allowed
PITCHING_H	Hits allowed
PITCHING_HR	Homeruns allowed
PITCHING_SO	Strikeouts by pitchers

A detailed Exploratory Data Analysis (EDA) section is developed to better understand the characteristics and properties of the variables. The EDA section endeavors to understand the distribution and shape of each variable provided, identify outliers and missing values, and understand correlation among the predictor variables as well as correlation with the response variable.

The deeper understanding gleaned from the EDA phase informs the subsequent data preparation and transformation process. The data preparation step attempts to optimize the inputs into the regression models by addressing predictor variables with (1) high collinearity, (2) sufficiently large numbers of missing values effectively rendering them unusable, (3) values identified as outliers deemed implausible based on historical baseball statistics, and (4) the creation of new predictor variables based on existing. Various techniques for imputing missing values are also explored and compared, resulting in the usage of Random Forest imputation as the method used to address missing values in the dataset.

The modeling building phase builds four models using different combinations of variables and selection approaches. Model 1 employs backward stepwise variable selection, purposely limited to the provided predictor variables or those created specifically to address collinearity from among the base variables. Model 2 starts with the derived batting variable “Total Bases” as a minimum predictor and applies forward stepwise selection. Model 3 uses a derived pitching statistic called “Walks plus Hits per Game Played” or WHGP and also applies forward stepwise variable selection. Finally, Model 4 uses a Sabmetric statistic called Base Runs or BsR to determine an optimal regression model through bi-directional stepwise variable selection. Common to all model building is the creation and analysis of summary statistics and diagnostics to support the next phase -- final model selection.

The project concludes with the selection of the best model from among the four models based on AIC, Adjusted R-squared, and predicted win accuracy against the training dataset.

Objective and Requirements

We are to build a multiple linear regression model on the training data to predict the number of wins for the team. Only the variables given (or variables that can be derived from the variables provided) will be used. The variables selections to be included in the model(s) will be done manually.

Data

For reproducibility of the results, we will load the data from Github repository. We will also remove the prefix of “TEAM” from the predictors variables to reduce cluttering in our plots and tables output.

Team

These are the members of the team that collaborated on this effort:
Sharon Morris
Brian Kreis
MichaelnD'Acampora
Keith Folsom
Valerie Briot

Contents

Data Exploration

The following variables comprise the data set. The "INDEX" variable is a unique identifier for the row and will have no bearing on the model and will be ignore. The variable "TARGET_WIN" is the variable of interest, the response variable that we are planning to predict via the model. This variable is of type count (continuous without fractional numbers). The remaining 15 variables are predictor variables that would possibly be selected when building the model. All the predictor variables are of type count (continuous without fractional numbers).

We have grouped the variables by categories; to identify the baseball statistic the variable represents (Batting, Fielding, Pitching, …).

Variable Name	**Definition**	**Theoretical Effect**	**Category**	**Variable Type**	Data Type
INDEX	Identification Variable	None	Identifier		
TARGET_WINS	Number of wins		Result	Response	Count
BATTING_H	Base Hits by batters (1B,2B,3B,HR)	Positive	Batting	Predictor	Count
BATTING_2B	Doubles by batters (2B)	Positive	Batting	Predictor	Count
BATTING_3B	Triples by batters (3B)	Positive	Batting	Predictor	Count
BATTING_HR	Homeruns by batters (4B)	Positive	Batting	Predictor	Count
BATTING_BB	Walks by batters	Positive	Batting	Predictor	Count
BATTING_HBP	Batters hit by pitch	Positive	Batting	Predictor	Count
BATTING_SO	Batters hit by pitch	Negative	Batting	Predictor	Count
BASERUN_SB	Stolen bases	Positive	Baserunning	Predictor	Count
BASERUN_CS	Caught stealing	Negative	Baserunning	Predictor	Count
FIELDING_E	Errors	Negative	Fielding	Predictor	Count
FIELDING_DP	Double Plays	Positive	Fielding	Predictor	Count
PITCHING_BB	Walks allowed	Negative	Pitching	Predictor	Count
PITCHING_H	Hits allowed	Negative	Pitching	Predictor	Count
PITCHING_HR	Homeruns allowed	Negative	Pitching	Predictor	Count
PITCHING_SO	Strikeouts by pitchers	Positive	Pitching	Predictor	Count

For each of these variables (excluding the "INDEX"), we will perform Exploratory Data Analysis (or EDA) to understand the characteristics of the data prior to modeling. Using the "Describe" function from the "psych" package, we will compute the major descriptive statistics for each variable (Mean, Median, Standard Deviation, …). For skewness and Kurtosis, we will use the type 3 method. The results have been summarized in a table (see next page).

	n	mean	sd	median	min	max	skew	kurtosis	se	IQR	Q0.1	Q0.25	Q0.75	Q0.9	missing	ratio
TARGET_WINS	2276	80.79086	15.75215	82.0	0	146	-0.3987232	1.0274757	0.3301823	21.00	61.0	71.0	92.00	99.5	0	0.0000
BATTING_H	2276	1469.26977	144.59120	1454.0	891	2554	1.5713335	7.2785261	3.0307891	154.25	1315.0	1383.0	1537.25	1635.5	0	0.0000
BATTING_2B	2276	241.24692	46.80141	238.0	69	458	0.2151018	0.0061609	0.9810087	65.00	182.0	208.0	273.00	303.0	0	0.0000
BATTING_3B	2276	55.25000	27.93856	47.0	0	223	1.1094652	1.5032418	0.5856226	38.00	27.0	34.0	72.00	96.0	0	0.0000
BATTING_HR	2276	99.61204	60.54687	102.0	0	264	0.1860421	-0.9631189	1.2691285	105.00	20.0	42.0	147.00	179.5	0	0.0000
BATTING_BB	2276	501.55888	122.67086	512.0	0	878	-1.0257599	2.1828544	2.5713150	129.00	363.5	451.0	580.00	635.0	0	0.0000
BATTING_SO	2174	735.60534	248.52642	750.0	0	1399	-0.2978001	-0.3207992	5.3301912	382.00	421.0	548.0	930.00	1049.0	102	4.4815
BASERUN_SB	2145	124.76177	87.79117	101.0	0	697	1.9724140	5.4896754	1.8955584	90.00	44.0	66.0	156.00	231.0	131	5.7557
BASERUN_CS	1504	52.80386	22.95634	49.0	0	201	1.9762180	7.6203818	0.5919414	24.00	30.0	38.0	62.00	77.0	772	33.9192
BATTING_HBP	191	59.35602	12.96712	58.0	29	95	0.3185754	-0.1119828	0.9382681	16.50	44.0	50.5	67.00	76.0	2085	91.6081
PITCHING_H	2276	1779.21046	1406.84293	1518.0	1137	30132	10.3295111	141.8396985	29.4889618	263.50	1356.0	1419.0	1682.50	2057.5	0	0.0000
PITCHING_HR	2276	105.69859	61.29875	107.0	0	343	0.2877877	-0.6046311	1.2848886	100.00	25.0	50.0	150.00	187.0	0	0.0000
PITCHING_BB	2276	553.00791	166.35736	536.5	0	3645	6.7438995	96.9676398	3.4870317	135.00	417.5	476.0	611.00	693.5	0	0.0000
PITCHING_SO	2174	817.73045	553.08503	813.5	0	19278	22.1745535	671.1891292	11.8621151	353.00	490.0	615.0	968.00	1095.0	102	4.4815
FIELDING_E	2276	246.48067	227.77097	159.0	65	1898	2.9904656	10.9702717	4.7743279	122.25	109.0	127.0	249.25	542.0	0	0.0000
FIELDING_DP	1990	146.38794	26.22639	149.0	52	228	-0.3889390	0.1817397	0.5879114	33.00	109.0	131.0	164.00	178.0	286	12.5659

From a first cursory glance at the results, we notice that the following variables have missing values; TEAM_BATTING_SO, TEAM_BASERUN_SB, TEAM_BASERUN_CS, TEAM_BATTING_HBP, TEAM_PITCHING_SO, and TEAM_FIELDING_DP.

Based on the Skewness and Kurtosis coefficients, some of the variables appears to have moderate skewness and kurtosis (TEAM_BATTING_H, TEAM_BASERUN_SB, TEAM_BASERUN_CS, …) and some have significant skewness and kurtosis (TEAM_PITCHING_H, TEAM_PITCHING_BB, TEAM_PITCHING_SO, TEAM_FIELDING_E). These may denote asymmetric distribution and heavy tails and therefore probably the persistence of outliers. We will further analyze these possibilities with histograms and boxplots.

Finally, quite a few variables have minimum of zero, one may question whether these are genuine values; for example, TEAM_BATTING_SO; it is unlikely that a team in a given year had no batter strike out. Further analysis on each individual variable will be completed.

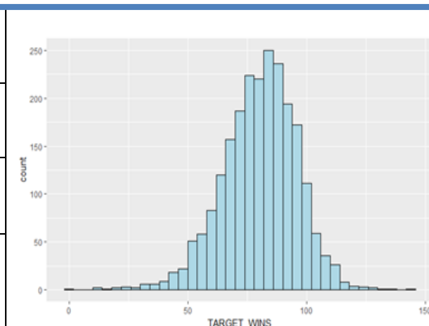
Linear regression models do not perform well with predictor with skewed distributions, outliers, and missing data. We will need to address these prior to building our models.

We will continue to explore the individual variables for further insights into the data.

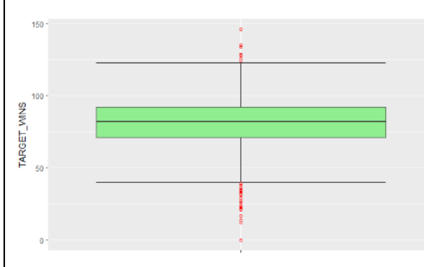
TARGET_WINS

This is our response variable. As with all the predictors, this is a continuous "count" variable.

n	**mean**	**sd**	**min**	**max**	**median**	**IQR**	**Q.25**	**Q.75**
2276	80.79086	15.75215	0	146	82	21	71	9
Skew	-0.3987232	**Kurtosis**	1.027476			**Msng Ratio**	0	
Outliers		146, 135, 134, 129, 33, 32, 31, 29, 27, 26, 24, 23, 22, 21, 17, 14, 12, 0						



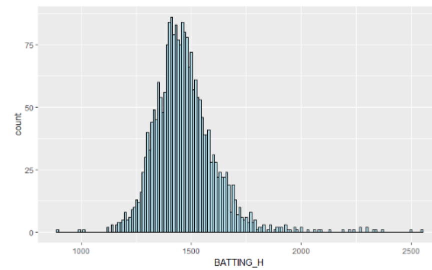
From the histogram and the boxplot, the distribution the distribution for the response variable distribution is fairly symmetrical. This confirmed the insight we had by looking at the skewness and kurtosis coefficients. When looking at the minimum and maximum for this variable, we had questions regarding the validity of the data. A value of 0 for minimum, would indicate that we have team(s) that did not win any games in a given season. This would be very unlikely. Also, a maximum 146 win a season (out of 162 since the historical data have been adjusted to match the performance of a 162 games season) is highly unlikely. We consulted the baseball almanac and were able to confirm that historically (adjusted for 162 games season) the minimum and maximum should be 22 and 124 respectively. We feel very comfortable therefore removing records with values outside of this range. We have no missing values for this variable.



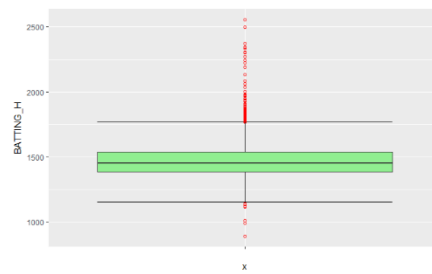
BATTING_H

This is one of our 16 predictor variables. It is a continuous “count” variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
2276	1469.27	144.5912	891	2554	1454	154.25	1383	1537.25
Skew	1.571333	**Kurtosis**	7.278526			**Msng Ratio**	0	
Outliers		2554, 2496, 2372, 2343, 2333, 2305, 2300, 2273, 2270, 2241, 2222, 2192, 2132, 2086						



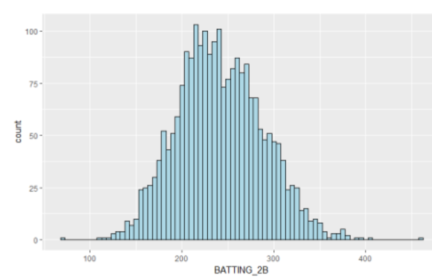
From the histogram and the boxplot, the distribution for the response variable distribution right-skewed with heavy tail. This is confirmed the insight we had by looking at the skewness and kurtosis coefficients. This variable has a presence of outliers as indicated by the boxplot and histogram. Because of the large presence of outliers, we had questions regarding the validity of the data. We consulted the baseball almanac and were able to confirm that historically (adjusted for 162 games season) the maximum should be 1876. We feel very comfortable therefore removing records with values greater than this maximum. We have no missing values for this variable.



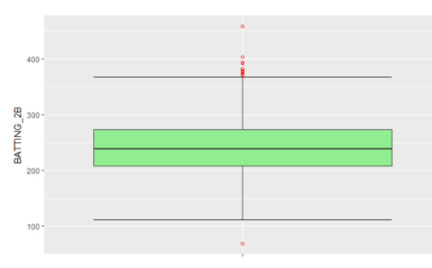
BATTING_2B

This is one of our 16 predictor variables. It is a continuous “count” variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
2276	241.2469	46.80141	69	458	238	65	208	273
Skew	0.2151018	**Kurtosis**	0.0061609			**Msng Ratio**	0	
Outliers		458, 403, 393, 392, 382, 69						



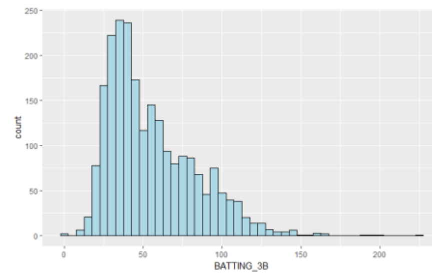
From the histogram and the boxplot, we can see that the distribution for this variable is fairly symmetrical. This is supported by the fact the mean and median are closed together in value and by the fairly low skewness and kurtosis coefficients. From the plots and our calculations, we can clearly see some outliers (particularly for high values). This made us question the validity of the data. We consulted the baseball almanac and were able to confirm that historically the minimum should be 116 (adjusted for 162 games season) and the maximum should be 376. We would feel comfortable removing the records with values outside this range. We have no missing values for this variable.



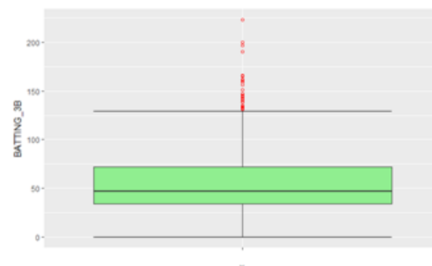
BATTING_3B

This is one of our 16 predictor variables. It is a continuous "count" variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
2276	55.25	27.93856	0	223	47	38	34	72
Skew	1.109465	**Kurtosis**	1.503242			**Msng Ratio**	0	
Outliers		223, 200, 197, 190, 166, 165, 162, 161, 159, 156, 151, 147, 145, 144, 143, 142, 140						



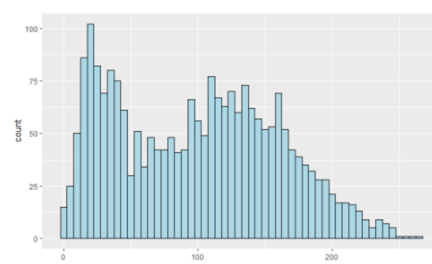
From the histogram and the boxplot, we can see that the distribution for this variable is right-skewed. This is supported by the fact the median is much larger than the mean and by the moderate skewness and kurtosis coefficients. From the plots and our calculations, we can clearly see some outliers in high values also the minimum for this variable is zero which seems very unlikely. This would mean that a team in a given year would have not hit any triples. This made us question the validity of the data. We consulted the baseball almanac and were able to confirm that historically the minimum should be 11 and the maximum should be 153. We would feel comfortable removing the records with values outside this range. We should note that a minimum of zero was not considered an outlier. There are no missing values for the variable.



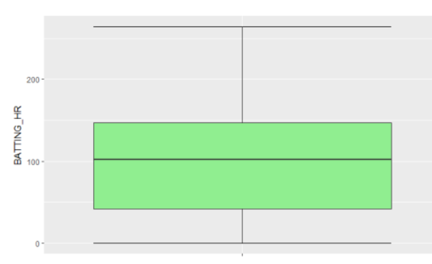
BATTING_HR

This is one of our 16 predictor variables. It is a continuous "count" variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
2276	99.61204	60.54687	0	264	102	105	42	147
Skew	0.1860421	**Kurtosis**	-0.9631189			**Msng Ratio**	0	
Outliers		None						



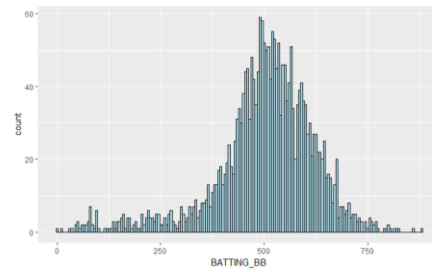
From the histogram, we can see that the distribution for this variable is bi-modal with a slight right-skewness. From the plots and our calculations, we do not detect any outliers. The minimum for this variable is zero which seems very unlikely. This would mean that a team in a given year would not have hit any home run. This made us question the validity of the data. We consulted the baseball almanac and were not able to correct any possible errors in the data. We should note that a minimum of zero was not considered an outlier. There is no missing value for the variable.



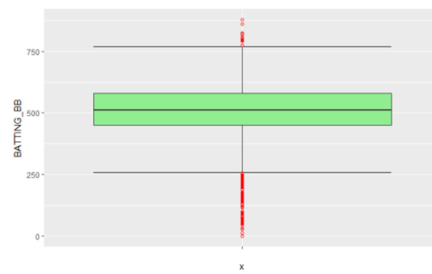
BATTING_BB

This is one of our 16 predictor variables. It is a continuous “count” variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
2276	501.5589	122.6709	0	878	512	129	451	580
Skew	-1.02576	**Kurtosis**	2.182854			**Msng Ratio**	0	
Outliers		878, 128, 125, 119, 116, 102, 97, 95, 94, 93, 89, 84, 83, 81, 79, 78, 74, 73, 72, 71, 65, 64, 59, 58, 56, 52, 50, 45, 34, 29, 12, 0						



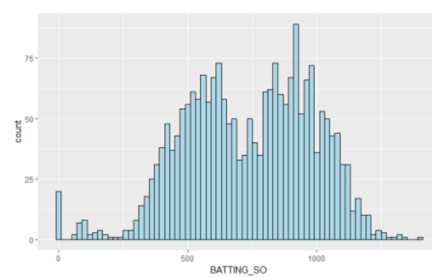
From the histogram and the boxplot, we can see that the distribution for this variable is left-skewed. This is supported by the moderate skewness and kurtosis coefficients. From the plots and our calculations, we can clearly see some outliers particularly in the low values also the minimum for this variable is zero which seems very unlikely. This would mean that a team in a given year would have no base-on-balls. This made us question the validity of the data. We consulted the baseball almanac and were able to confirm that historically the minimum should be 292 and the maximum should be 879 (both numbers were adjusted for 162 games season). We would feel comfortable removing the records with values outside this range. There is no missing value for the variable.



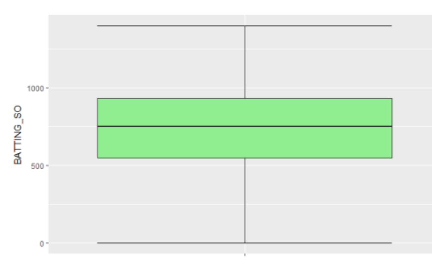
BATTING_SO

This is one of our 16 predictor variables. It is a continuous “count” variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
2174	735.6053	248.5264	0	1399	750	382	548	930
Skew	-0.2978001	**Kurtosis**	-0.3207992			**Msng Ratio**	4.5	
Outliers		None						



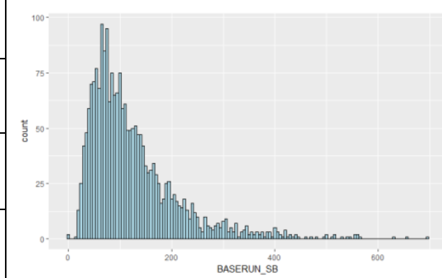
From the histogram and the boxplot, we can see that the distribution for this variable show a tendency for bi-modal with a slight left-skewness. This is supported the moderate skewness and kurtosis coefficients. From the plots and our calculations, are not showing any outliers. We are seeing a minimum of zero which made us question the validity of the data. We consulted the baseball almanac and were able to confirm that historically the minimum should be 326 and the maximum should be 1535. We would feel comfortable removing the records with values outside this range. We should note that a minimum of zero was not considered an outlier. About 4.5% of the data is missing for this variable. We will need to account for this.



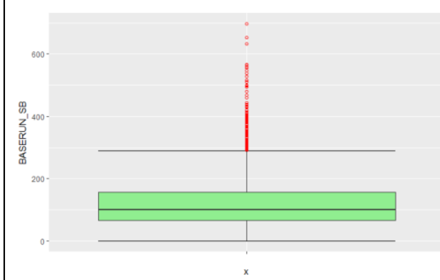
BASERUN_SB

This is one of our 16 predictor variables. It is a continuous "count" variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
2145	124.7618	87.79117	0	697	101	90	66	256
Skew	1.972414	**Kurtosis**	5.489675			**Msng Ratio**	5.8	
Outliers		697, 654, 632, 567, 562, 558, 555, 554, 547, 538, 529, 517, 511, 500, 499, 494, 481, 468, 460, 444, 439, 438, 433, 430, 429, 427, 420, 419, 414, 410, 406, 403, 401, 400, 399, 398, 394, 392						



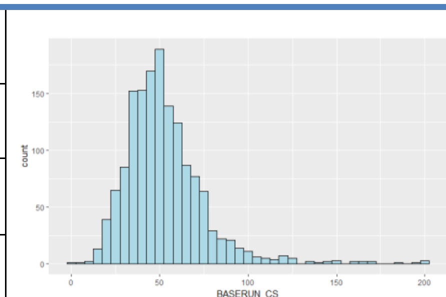
From the histogram and the boxplot, we can see that the distribution for this variable is right-skewed. This is supported by the skewness and kurtosis coefficients. From the plots and our calculations, we can clearly see some outliers in the high values also the minimum for this variable is zero which seems very unlikely. This would mean that at least one team in a given year would not have stolen one base. This made us question the validity of the data. We consulted the baseball almanac and were not able to eliminate values based on historical information. About 5.8% of the values for this variable are missing. We will need to address these missing values.



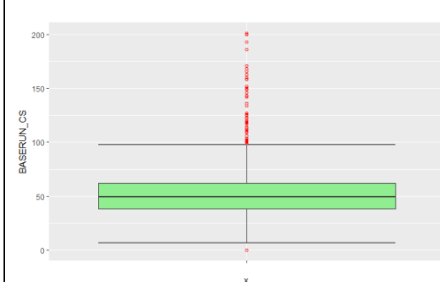
BASERUN_CS

This is one of our 16 predictor variables. It is a continuous "count" variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
1504	52.80386	22.95634	0	201	49	24	38	62
Skew	1.976218	**Kurtosis**	7.620382			**Msng Ratio**	34	
Outliers		201, 200, 193, 186, 171, 168, 166, 163, 160, 158, 152, 150, 149, 146, 143, 142, 136, 134, 127, 126, 125, 123, 122						



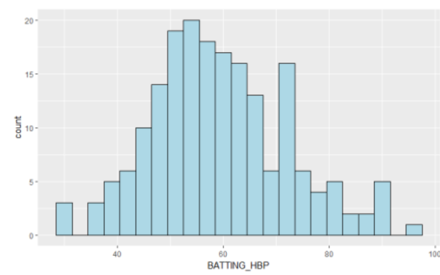
From the histogram and the boxplot, we can see that the distribution for this variable is right-skewed. This is supported by the skewness and kurtosis coefficients. From the plots and our calculations, we can clearly see some outliers particularly in the high values also the minimum for this variable is zero which seems very unlikely. This would mean that at least one team in a given year would not have been caught stealing a base. This made us question the validity of the data. We consulted the baseball almanac and were not able to eliminate values based on historical information. About 34% of the values for this variable are missing. We will need to address these values.



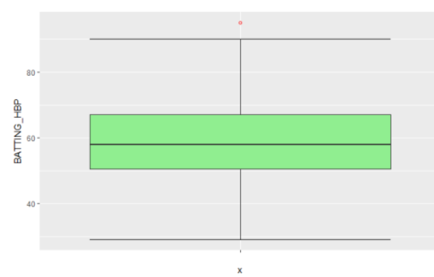
BATTING_HBP

This is one of our 16 predictor variables. It is a continuous “count” variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
191	59.35602	12.96712	29	95	58	16.5	50.5	67
Skew	0.3185754	**Kurtosis**	-0.1119828			**Msng Ratio**		91.6
Outliers		None						



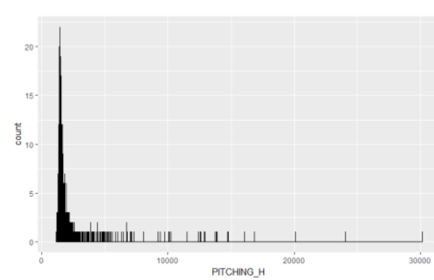
From the histogram and the boxplot, we can see that the distribution for this variable is fairly symmetrical. This is supported by the skewness and kurtosis coefficients. From the boxplot, we can see one outlier however this outlier is within 3 standard deviation from the mean as per our calculations. About 91.6% of the values for this variable is missing. This needs to be addressed.



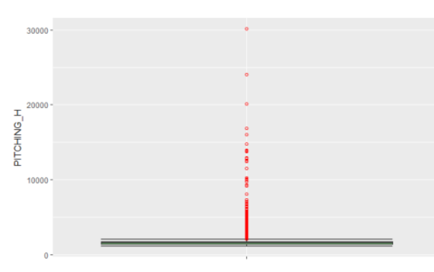
PITCHING_H

This is one of our 16 predictor variables. It is a continuous “count” variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
2276	1779.21	1406.843	1137	30132	1518	263.5	1419	1682.5
Skew	10.32951	**Kurtosis**	141.8397			**Msng Ratio**		0
Outliers		None						



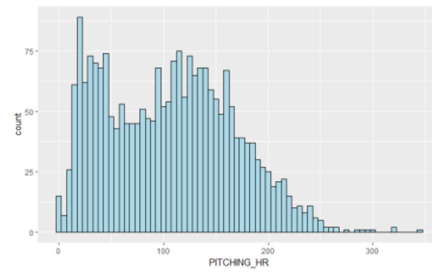
From the histogram and the boxplot, we can see that the distribution for this variable is highly right-skewed. This is supported by the significant skewness and kurtosis coefficients. From the plots and our calculations, we can clearly see some outliers in the high values. We have been able to confirm that the maximum number of hit for a team was 1876. This would make very unlikely that a pitching staff would have surrender more than 3,000 hits. We will remove records with values higher than this derived limit. There are no missing values for this variable.



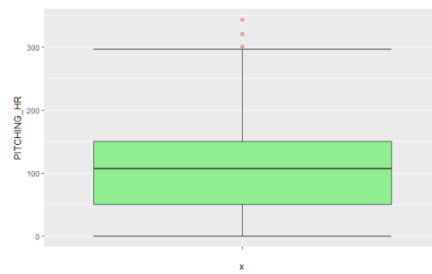
PITCHING_HR

This is one of our 16 predictor variables. It is a continuous “count” variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
2276	105.6986	61.29875	0	343	107	100	50	150
Skew	0.2877877	**Kurtosis**	-0.6046311			**Msng Ratio**		0
Outliers		343, 320, 301, 297, 291						



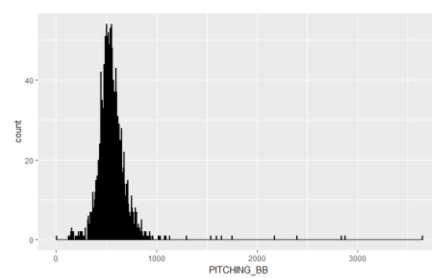
From the histogram and the boxplot, we can see that the distribution for this variable is slightly right-skewed. This is supported by the skewness and kurtosis coefficients. From the plots and our calculations, we can clearly see some outliers in the high values. We also have a minimum of 0 which seems unlikely. This would mean that a team would have not have given up a home run in a given year. We consulted the baseball almanac and were able to confirm that historically (adjusted for 162 games season) the maximum should be 258. We would consider removing any records with a value in this variable higher than this maximum. There is no missing value for this variable



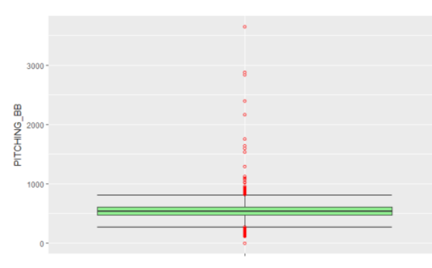
PITCHING_BB

This is one of our 16 predictor variables. It is a continuous “count” variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
2276	553.0079	166.3574	0	3645	536.5	135	476	611
Skew	6.7439	**Kurtosis**	96.96764			**Msng Ratio**		0
Outliers		3645, 2876, 2840, 2396, 2169, 1750, 1643, 1594, 1539, 1296, 1123, 1090, 1076, 0,						



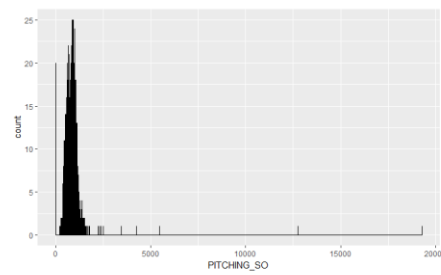
From the histogram and the boxplot, we can see that the distribution for this variable is right-skewed. This is supported by the skewness and kurtosis coefficients. From the plots and our calculations, we can clearly see extreme outliers. We also have a minimum of 0 which seems unlikely. This would mean that a team would have not have allowed a base-on-ball in a given year. We were not able to find evidence to discard these outliers. There is no missing value for this variable.



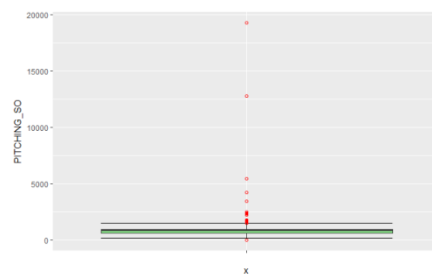
PITCHING_SO

This is one of our 16 predictor variables. It is a continuous "count" variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
2174	817.7305	553.085	0	19278	813.5	353	615	968
Skew	22.17455	**Kurtosis**	671.1891			**Msng Ratio**		4.5
Outliers		19278, 12758, 5456, 4224, 3450, 2492						



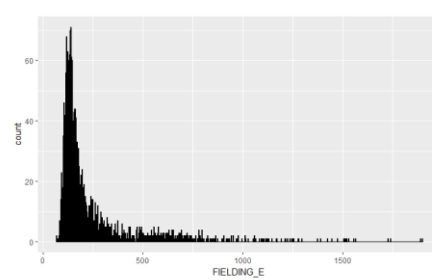
From the histogram and the boxplot, we can see that the distribution for this variable is right-skewed with extreme outliers in the right tail. This is supported by the skewness and kurtosis coefficients. From the plots and our calculations, we can clearly see extreme outliers. We also have a minimum of 0 which seems unlikely. This would mean that a team would have not have pitched a strike-out in a given year. We were able to confirm after consulting the baseball almanac that the maximum for this value should be 1450. We would therefore be confident in removing any records with a value higher than this maximum. We were not able to disprove the minimum of zero as unlikely as it appears to be. About 4.5% of the data is missing for this variable. This would need to be addressed.



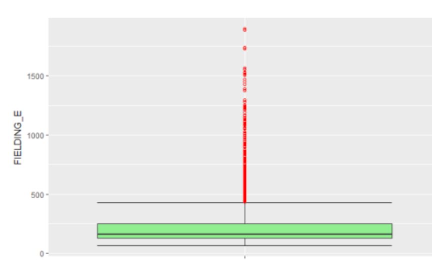
FIELDING_E

This is one of our 16 predictor variables. It is a continuous "count" variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
2276	246.4807	227.771	65	1898	159	122.25	127	249.25
Skew	2.990466	**Kurtosis**	10.97027			**Msng Ratio**		0
Outliers		1898, 1890, 1740, 1728, 1567, 1553, 1531, 1522, 1515, 1512, 1506, 1473, 1447, 1426						



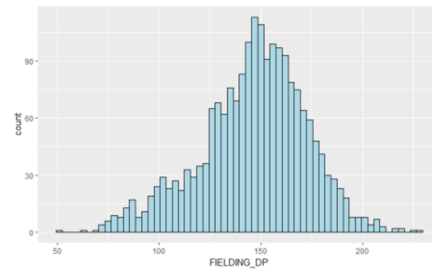
From the histogram and the boxplot, we can see that the distribution for this variable is highly right-skewed and have extreme outliers in the right tail. This is supported by the skewness and kurtosis coefficients. It seemed very unlikely that a team will commit that many errors in a year. There is no missing data for this variable.



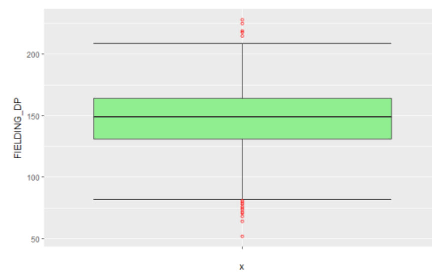
FIELDING_DP

This is one of our 16 predictor variables. It is a continuous "count" variable.

n	mean	sd	min	max	median	IQR	Q.25	Q.75
1990	146.3879	26.22639	52	228	149	33	131	164
Skew	-0.388939	**Kurtosis**	0.1817397			**Msng Ratio**		12.6
Outliers		None						

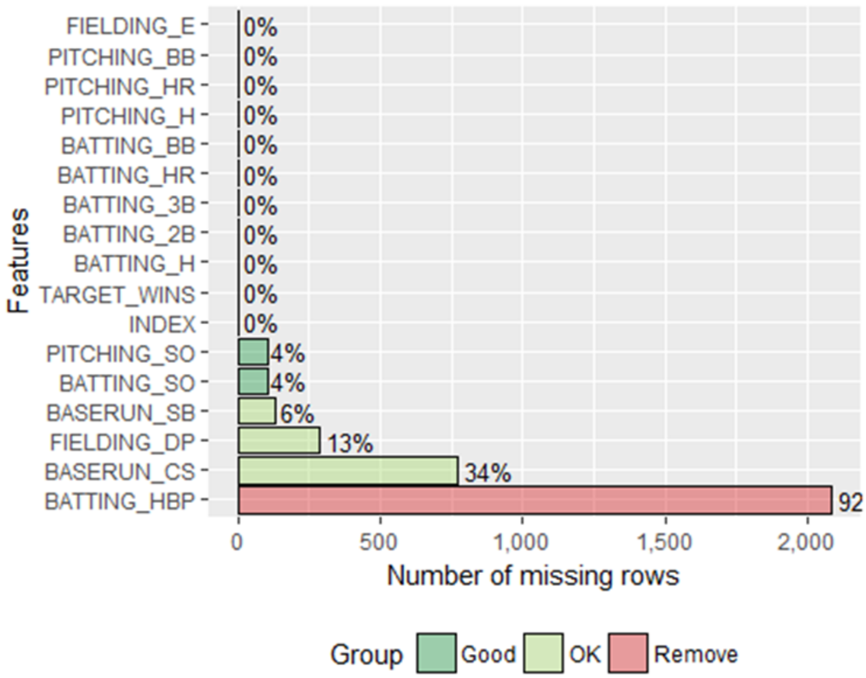


From the histogram and the boxplot, we can see that the distribution for this variable is left-skewed. This is supported by the skewness and kurtosis coefficients. From the boxplot and our calculations there are some outliers in the data. About 12.6% of the data for this variable is missing and would need to be addressed prior to building a model.



Missing Values

As we encountered in our exploratory data analysis, we have a few variables with missing values. A strategy will be devised to handle these prior to building our model.



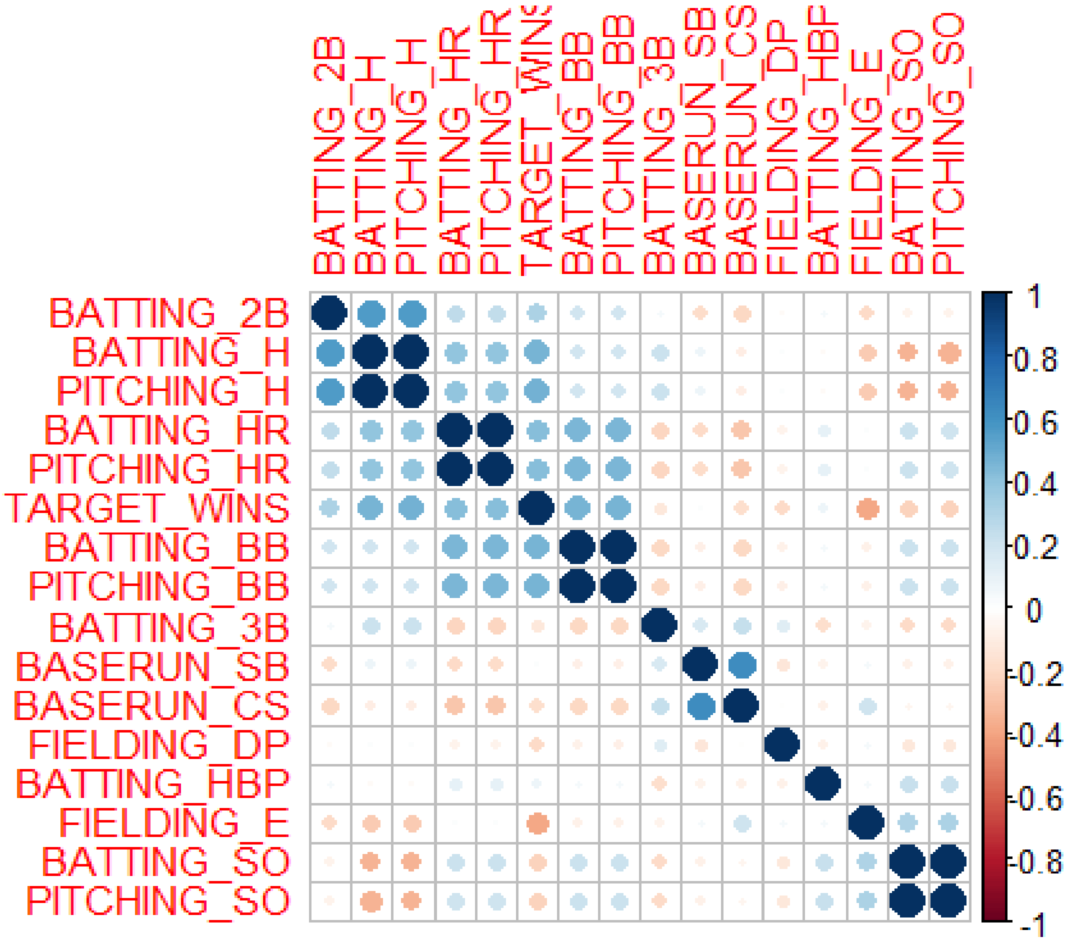
There are missing values for the variables PITCHING_SO, BATTING_SO, BASERUN_SB, FIELDING_DP, BASERUN_CS, and BATTING_HBP.

The variable BATTING_HBP has the most missing values at 92% missing or 2085 out of 2276 observations and, as a result, we may consider excluding the variable.

The variable BASERUN_CS has the next most missing values at 34% missing or 772 out of 2276 observations.

(The graph was produced by the plot_missing function from DataExplorer package).

Correlation between variables



From the correlation matrix, we can see that the problem variable (BATTING_HBP), with 92% of missing data, does not show correlation with our response variable (TARGET_WINS).

Also, we have very positive correlation between BATTING_H and PITCHING_H, between BATTING_BB and PITCHING_BB, between BATTING_HR and PITCHING_HR, and between BATTING_SO and PITCHING_SO. This makes sense intuitively since these measures are really measuring the same thing from offense (batting) or defense (pitching) point of view.

Furthermore, we have positive correlation between BATTING_H and BATTING_2B and with BATTING_H and BATTING_HR and to a lesser degree with BATTING_3B. This is not surprising since BATTING_H should encompass the other hit batting statistics.

Finally, BASERUN_SC is strongly correlated with BASERUN_SB and is missing about 34% of the data.

VARIABLE	CORRELATION WITH WINNING
TEAM_BATTING_H	0.3887675
TEAM_BATTING_2B	0.2891036
TEAM_BATTING_3B	0.1426084
TEAM_BATTING_HR	0.1761532
TEAM_BATTING_BB	0.2325599
TEAM_BATTING_SO	-0.0317507
TEAM_BASERUN_SB	0.1351389
TEAM_BASERUN_CS	0.0224041
TEAM_BATTING_HBP	0.0735042
TEAM_PITCHING_H	-0.1099371
TEAM_PITCHING_HR	0.1890137
TEAM_PITCHING_BB	0.1241745
TEAM_PITCHING_SO	-0.0784361
TEAM_FIELDING_E	-0.1764848
TEAM_FIELDING_DP	-0.0348506

Data Transformation

The following data transformations will be performed irrespective to the model we are building. Additional transformation may be added for an individual model.

Removal of predictor variable due to collinearity and/or Missing values

BATTING_HBP:
A relatively small correlation with our response variable and is missing over 90% of the data. We will therefore remove this variable from our analysis and not incorporate it in our model building.

BASERUN_SC:
Since this variable is strongly correlated to BASERUN_SB and is missing about 34% of data, we will remove this variable from our analysis and not incorporate it in our model building.

BATTING_H:
We will replace this variable with BATTING_1B, derived as follows:
BATTING_1B = BATTING_H - (BATTING_2B + BATTING_3B + BATTING_HR)
This transformation will be done once some outliers for BATTING_H have been handled and missing values have been imputed for the dataset.

Removal of Egregious outliers from data

With our research in the Baseball Almanac and with Subject Matter Expertise, we will remove the following outliers from the data set.

TARGET_WINS:
We will remove records with values outside of researched historical range of [22,124]

BATTING_H:
We will remove records with values higher than researched maximum historical value of 1876

BATTING_2B:
We will remove records with values outside of researched historical range of [116,376]

BATTING_3B:
We will remove records with values outside of researched historical range of [11,153]

BATTING_BB:
We will remove records with values outside of researched historical range of [292,879]

BATTING_SO:
We will remove records with values outside of researched historical range of [326,1535]

PITCHING_HR:
We will remove records with values higher than researched maximum historical value of 258

PITCHING_SO: We will remove records with values higher than researched maximum historical value of 1450

PITCHING_H:
We will remove records with values higher than derived limit of 3,000, we have confirmed that the maximum value for BATTING_H was 1876 in season, hence, we can concluded that the number of hits pitched should not be greater than 3000.

Replacing remaining 0 values with NA

Based on our research and SME knowledge in our team, we will replace the remaining 0 values with NA and handle these as part of resolution for missing values.

We have removed 297 observations from our original training data set.

Addressing Skewness of Some Variables with Box-Cox

Some of our variables, most notably BATTING_3B, BASERUN_SB, PITCHING_H, PITCHING_BB, and FIELDING_E have pronounced skewness. We discussed whether we should transform these variables with Box Cox transformation. We had concerns on how imputed values for BASERUN_SB might be negatively impacted. However, BASERUN_SB to be inputted represents only 0.81% of the data and we would like to keep the interpretation of the model as simple as possible. As we build the model and evaluate them we may bring some transformations to address problems with the residuals.

Adding some additional predictors:

1. Replace BATTING_H with BATTING_1B

In our data, we have BATTING_H for total hits and individual values for double (BATTING_2B), triple (BATTING_3b), and Homerun (BATTING_HR). As we saw we have collinearity between BATTING_H and the over 3 batting statistics.

BATTING_1B = BATTING_H - (BATTING_2B - BATTING_3B - BATTING_HR)

2. Adding BATTING_TB

Total number of bases (BATTING_TB) is an additional measure that we feel is more representative as it gives a weight to each hit scored. Total number of bases takes into account the production underlying each hit giving a heavier weighting to doubles, triples and home runs.

BATTING_TB = BATTING_1B + 2xBATTING_2B + 3xBATTING_3B + 4xBATTING_HR

3. Adding Walk-Hit average, Pitched Strike-out to Walk ratio, and batted Walk to Shrike-out ratio

WHGP = (PITCHING_H + PITCHING_BB)/162

PITCHING_SO_BB = PITCHING_SO/PITCHING_BB

BATTING_BB_SO = BATTING_BB/BATTING_SO

4. Adding BsR

Base Runs (BsR) is a sabermetric stat created by David Smyth, to predict the number of runs a team would be expected to have scored based on the types of hits and number of walks that they had. We will estimate this measure.

BsR is calculated as follows:

$$BsR = \frac{A \cdot B}{B + C} + D$$

$$A = BATTING\_H + BATTING\_BB - BATTING\_HR$$

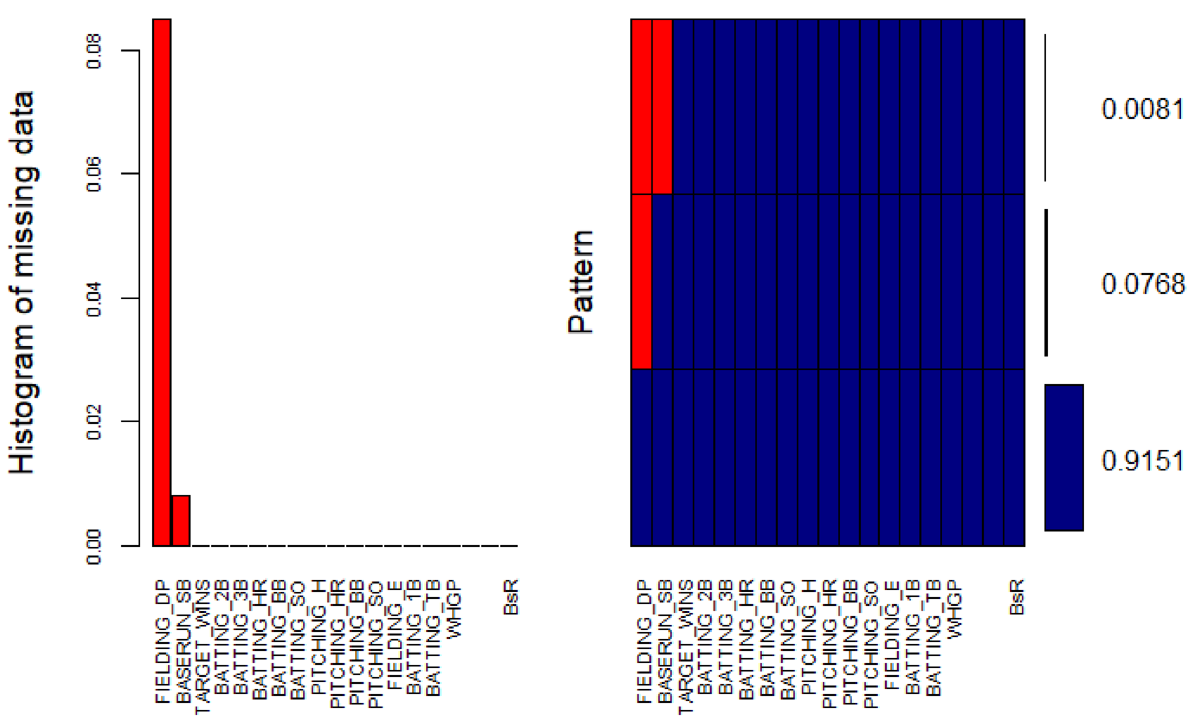
$$B = 1.4 \times (BATTING\_1B + 2 \times BATTING\_2B + 3 \times BATTING\_3B + 4 \times BATTING\_HR) - 0.6 \times BATTING\_H - 3 \times BATTING\_HR + 0.1 \times BATTING\_BB$$

$$\mathrm{C} = \mathrm{A} \cdot \mathrm{B} - \mathrm{BATTING\_H}$$

$$D = BATTING\_HR$$

Imputation of Missing Values

We will now address the missing values in the remaining predictors. The MICE and VIM packages were used to further our analysis of the missing values now that we have removed some outliers and the 2 predictor variables with the most missing values.



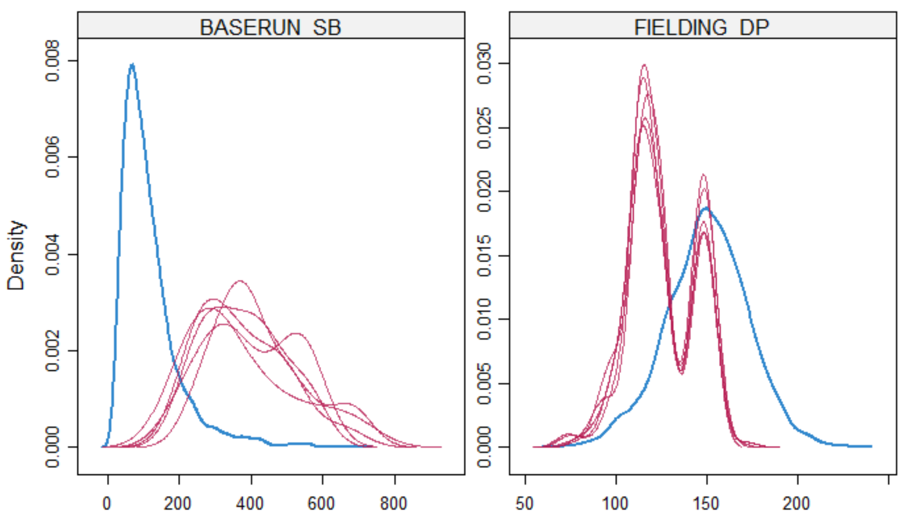
From this graph, we can clearly see that 91.5% of our data is complete, 7.7% is missing some values for predictor FIELDING_DP and remaining part of our data is missing both FIELDING_DP and BASERUN_SB. This is a much better picture as we should note that outlier removal and removal of some problematic predictors (BATTING_HBP and possibly BASERUN_CS) has improved the missing data picture. We will impute the remaining missing variables.

We selected 4 methods from the MICE packages to impute the missing values:

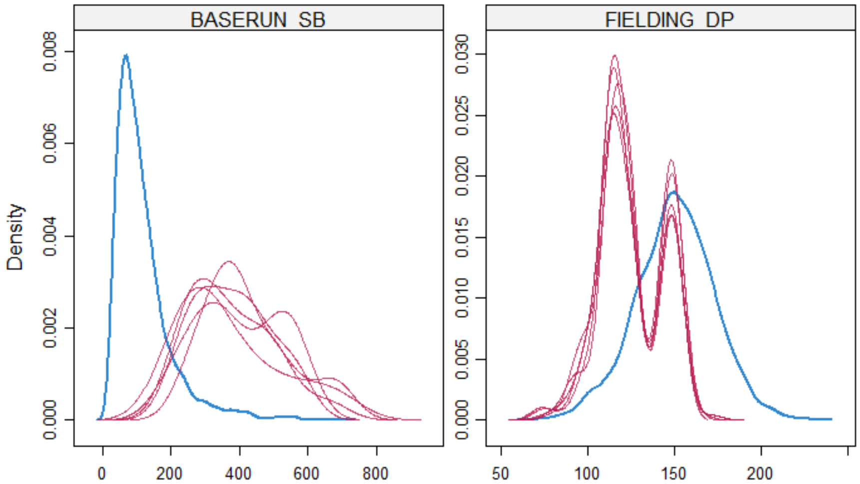
- Default method
- Predictive mean matching (pmm)
- Linear regression using bootstrap (norm.boot)
- Random forest imputations (rf)

In the density plots below, the imputed values are show in magenta and the values of the observed data are shown in blue.

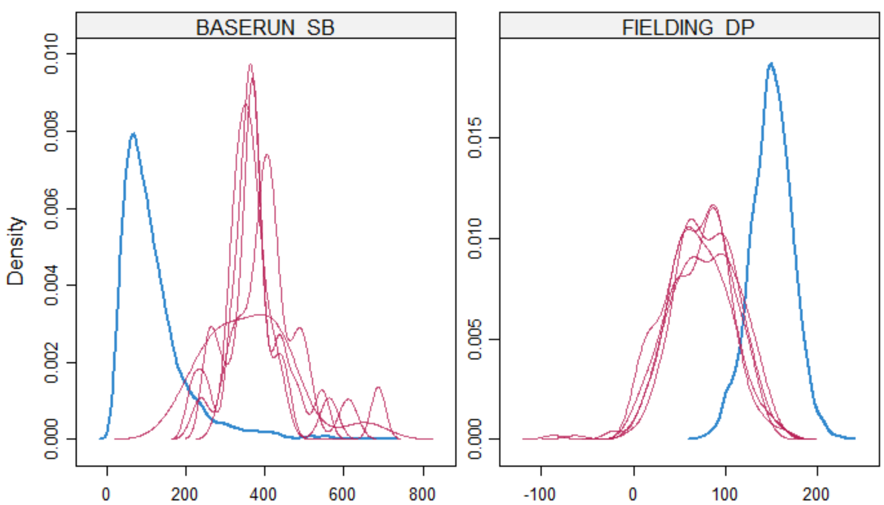
Default Method



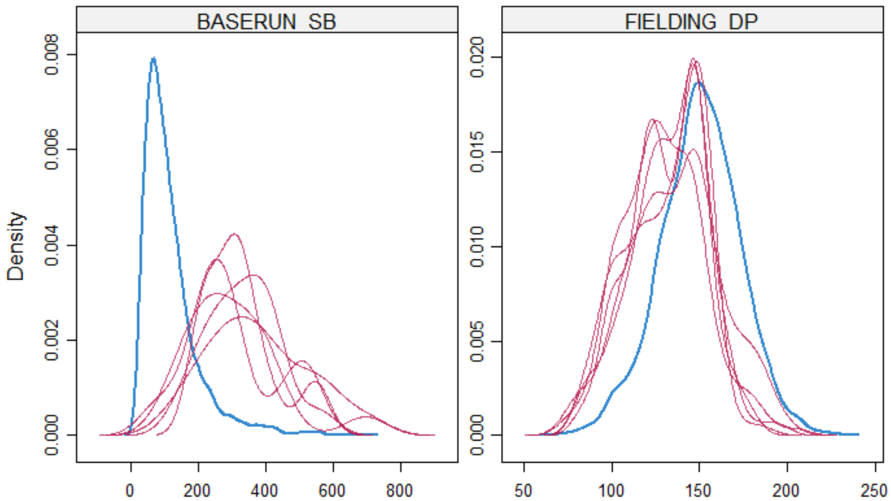
Predictive mean matching



Linear regression using bootstrap



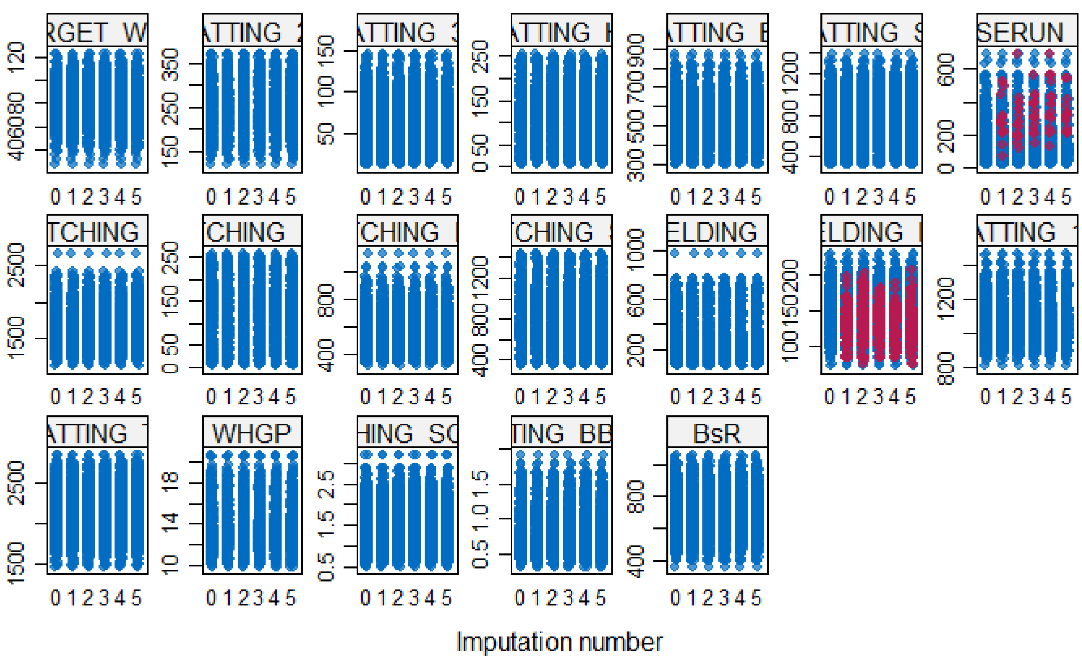
Random Forest Imputations



As we expected, imputed values for BASERUN_SB have been affected by the skewness of the data and the remaining presence of outliers. However, since only a minimal number of values for this variable need to be imputed (0.81%), we will proceed with these results.

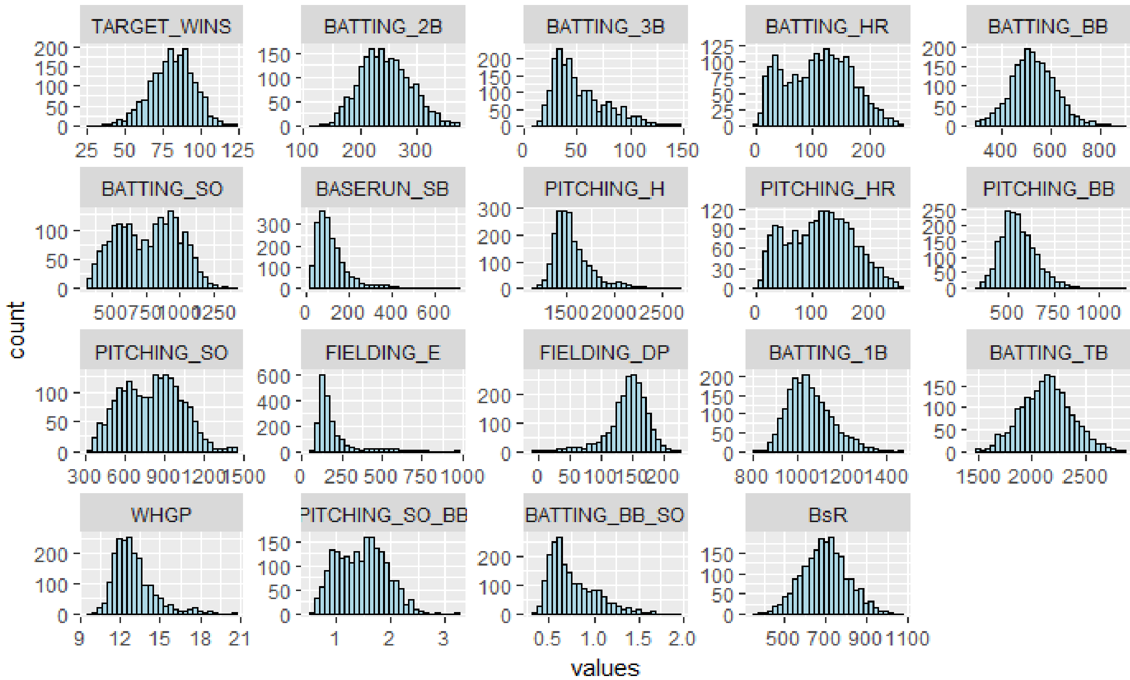
Based on the density plots, we will the Random Forest Imputation method (rf)

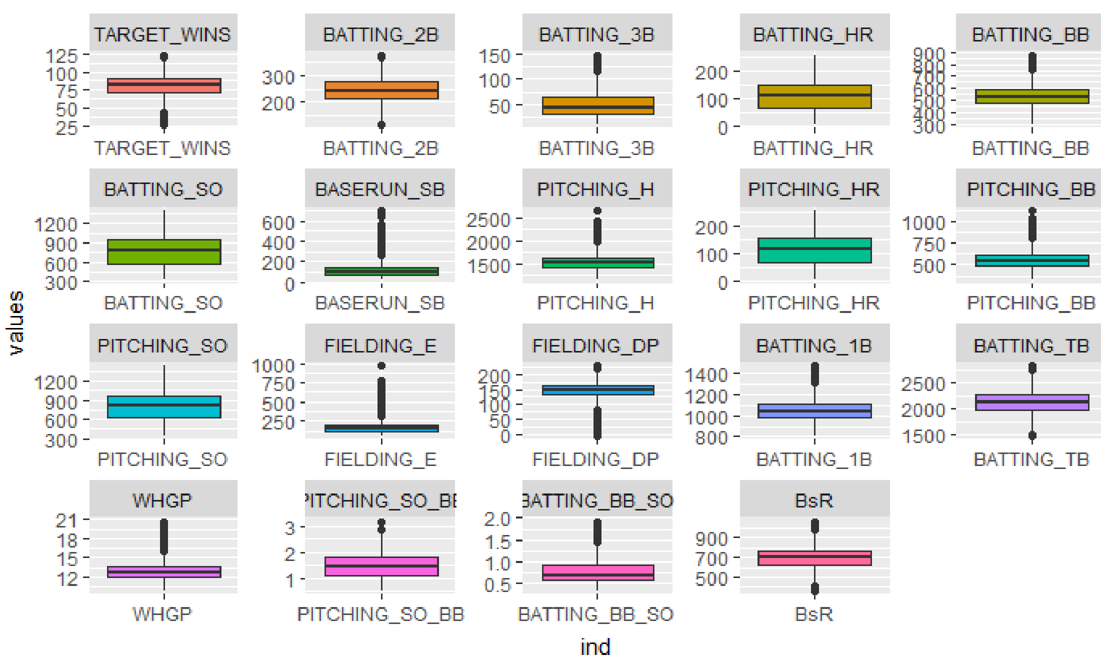
Strip-plot for imputed values using Random Forest method:



Transformation Recap

Now that we have completed the non-model specific transformation, we will quickly recap the exploration of our data set.





We have addressed some of the problems with the data, including egregious outliers and missing values and by addressing some extreme outliers with have reduced some of the skewness in our data. However, our data still shows some skewness. As we are building our models and refining them, we may have to transform further our variables possibly using box-cox method.

	n	mean	sd	median	mad	min	max	range	skew	kurtosis	se
TARGET_WINS	1979	80.9166246	13.8930852	82.0000000	14.8260000	27.0000000	123.000000	96.000000	-0.2310207	-0.0287605	0.3123027
BATTING_2B	1979	244.2299141	43.3020314	241.0000000	44.4780000	118.0000000	373.000000	255.000000	0.2074906	-0.3469930	0.9733866
BATTING_3B	1979	51.8120263	25.2529176	44.0000000	22.2390000	11.0000000	147.000000	136.000000	1.0052428	0.4490932	0.5676605
BATTING_HR	1979	109.5593734	56.1087335	112.0000000	65.2344000	4.0000000	257.000000	253.000000	0.0738565	-0.8386680	1.2612686
BATTING_BB	1979	527.5199596	88.1659777	524.0000000	87.4734000	294.0000000	878.000000	584.000000	0.1933220	0.2439120	1.9818835
BATTING_SO	1979	763.5234967	221.8707866	784.0000000	271.3158000	326.0000000	1399.000000	1073.000000	0.0011300	-1.0019470	4.9874347
BASERUN_SB	1979	119.8020502	86.7455815	97.0000000	57.8214000	18.0000000	697.000000	679.000000	2.2490360	7.1311821	1.9499544
PITCHING_H	1979	1547.6478019	196.4911043	1505.0000000	151.2252000	1137.0000000	2656.000000	1519.000000	1.4144464	2.6188413	4.4169247
PITCHING_HR	1979	113.6887317	56.4022561	115.0000000	63.7518000	4.0000000	257.000000	253.000000	0.0932354	-0.7918527	1.2678667
PITCHING_BB	1979	557.0454775	99.9976388	545.0000000	90.4386000	325.0000000	1123.000000	798.000000	0.8415029	1.4413100	2.2478475
PITCHING_SO	1979	799.7412835	216.6012672	811.0000000	247.5942000	345.0000000	1434.000000	1089.000000	0.1180996	-0.6304723	4.8689812
FIELDING_E	1979	192.3880748	124.0391869	149.0000000	47.4432000	65.0000000	965.000000	900.000000	2.4539554	6.0323368	2.7882776
FIELDING_DP	1979	143.9618167	31.9192014	148.0000000	25.2042000	-4.5710316	228.000000	232.571032	-1.2754889	2.8874291	0.7175119
BATTING_1B	1979	1055.0490147	96.9957528	1042.0000000	93.4038000	811.0000000	1464.000000	653.000000	0.6824080	0.4600018	2.1803681
BATTING_TB	1979	2137.1824154	227.0700252	2140.0000000	231.2856000	1478.0000000	2832.000000	1354.000000	0.0273203	-0.2066965	5.1043085
WHGP	1979	12.9919338	1.5345373	12.6975309	1.2080444	9.8395062	20.679012	10.839506	1.4267005	2.8561937	0.0344949
PITCHING_SO_BB	1979	1.4742367	0.4513862	1.4820031	0.5129036	0.5209040	3.185941	2.665037	0.2128807	-0.4938300	0.0101467
BATTING_BB_SO	1979	0.7536655	0.2633636	0.6747624	0.2212040	0.3136033	1.920981	1.607378	1.0767949	0.8070934	0.0059202
BsR	1979	697.7613106	108.2990387	696.2184591	104.2822422	360.3670605	1060.506614	700.139553	0.1252850	-0.0411831	2.4344547

Building Models

Model 1 - Base Variables

Model 1.1

For this model the dependent is TARGET_WINS and all predictors in the dataset are used for the initial model to, determine the importance of the predictors in predicting the dependent variable.

Our predictors are; PITCHING_H + PITCHING_HR + PITCHING_BB + PITCHING_SO + PITCHING_SO_BB + BATTING_2B + BATTING_3B + BATTING_HR + BATTING_BB + BATTING_SO + BATTING_1B + BATTING_TB + BATTING_BB_SO + FIELDING_E + FIELDING_DP + BASERUN_SB + BsR

We will use the backward regression technique. This regression technique is used to determine the best predictor variable by adding all predictors to the model. After the model with all of the predictors is created the Akaike Information Criterion (AIC) is reviewed. The AIC provide information about which predictors should be removed from the model to create the best fit.

Predictors with the highest numbers are removed and the model is executed again to determine whether the importance to the model of the remaining predictor variables. The process of removing variables is performed until removing variables will no longer lower the AIC.

The AIC considers the fit of the model and the number of parameters. All other things equal, the more predictor variables that are used in the model, the higher the AIC. The AIC penalizes when a model has more parameters, the number of parameters must be reduced to improve the model.

The magnitude of the AIC value is not of importance. Instead using the model with the lowest AIC value indicates the predictors that are the best fit.

The AIC value suggest removal of the first three variables (BSR, BATTING_1B and PRITCHING_SO) will drop the AIC to 9395 from 9393, thus improving the model.

Summary for model 1.1:

```
Residuals:
    Min      1Q  Median      3Q     Max 
-40.357  -6.938   0.012   6.944  47.467 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     35.120603   7.137530   4.921 9.35e-07 ***
PITCHING_H       0.071381   0.008675   8.228 3.41e-16 ***
PITCHING_HR     -0.191683   0.064976  -2.950 0.003215 ** 
PITCHING_BB     -0.149522   0.025908  -5.771 9.12e-09 ***
PITCHING_SO_BB  17.478624   3.035769   5.758 9.88e-09 ***
BATTING_2B      -0.081610   0.011716  -6.966 4.43e-12 ***
BATTING_3B       0.114012   0.018959   6.014 2.16e-09 ***
BATTING_HR       0.249909   0.067422   3.707 0.000216 ***
BATTING_BB       0.264966   0.029219   9.068  < 2e-16 ***
BATTING_SO      -0.069789   0.007478  -9.333  < 2e-16 ***
BATTING_1B      -0.039170   0.010263  -3.817 0.000139 ***
BATTING_BB_SO  -18.235017   3.486053  -5.231 1.87e-07 ***
FIELDING_E      -0.101922   0.004634 -21.993  < 2e-16 ***
FIELDING_DP     -0.130336   0.012425 -10.490  < 2e-16 ***
BASERUN_SB       0.068748   0.004754  14.460  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.69 on 1964 degrees of freedom
Multiple R-squared:  0.4122,	Adjusted R-squared:  0.408 
F-statistic: 98.38 on 14 and 1964 DF,  p-value: < 2.2e-16
```

We will now take this model and perform another backward regression.

Model 1.2

Our predictors are; PITCHING_H + PITCHING_HR + PITCHING_BB + PITCHING_SO_BB + BATTING_2B + BATTING_3B + BATTING_HR + BATTING_BB + BATTING_SO + BATTING_TB + BATTING_BB_SO + FIELDING_E + FIELDING_DP + BASERUN_SB

The AIC value suggest removal of the first variable (BATTING_2B) will drop the AIC to 9390 from 9392, thus improving the model.

Summary for model:

```
Residuals:
    Min      1Q  Median      3Q     Max 
-40.327  -6.948   0.009   6.929  47.394 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      35.659404   6.688631   5.331 1.09e-07 ***
PITCHING_H        0.072522   0.006890  10.526  < 2e-16 ***
PITCHING_HR      -0.192645   0.064809  -2.973  0.00299 ** 
PITCHING_BB      -0.152450   0.022099  -6.899 7.06e-12 ***
PITCHING_SO_BB   17.462216   3.034088   5.755 1.00e-08 ***
BATTING_3B        0.235903   0.021505  10.970  < 2e-16 ***
BATTING_HR        0.413415   0.071255   5.802 7.62e-09 ***
BATTING_BB        0.267951   0.025761  10.401  < 2e-16 ***
BATTING_SO       -0.069852   0.007470  -9.350  < 2e-16 ***
BATTING_TB       -0.041007   0.005782  -7.093 1.83e-12 ***
BATTING_BB_SO   -18.168351   3.471604  -5.233 1.84e-07 ***
FIELDING_E       -0.101961   0.004630 -22.023  < 2e-16 ***
FIELDING_DP      -0.130086   0.012368 -10.518  < 2e-16 ***
BASERUN_SB        0.068801   0.004747  14.493  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.69 on 1965 degrees of freedom
Multiple R-squared:  0.4122,	Adjusted R-squared:  0.4083 
F-statistic:   106 on 13 and 1965 DF,  p-value: < 2.2e-16
```

Model 1.3

This model excludes variable (BATTING_2B) based on the results of model 1.2.

The predictors are; PITCHING_H + PITCHING_HR + PITCHING_BB +PITCHING_SO_BB + BATTING_3B + BATTING_HR + BATTING_BB + BATTING_SO + BATTING_TB + BATTING_BB_SO + FIELDING_E + FIELDING_DP + BASERUN_SB

The output of model 1.3 indicates, the removal of any more predictor variables will not improve the model by lowering the AIC. Thus, this is the end of the process of using the AIC to identify the importance of the predictor variables.

R-squared value

The R-squared value of 0.41 indicates model 3 explains 41% of the variability around TOTAL_WINS.

The R-squared value ranges between 0%-100%, a higher R-squared value is desirable. A value of 0% indicates that 0% of the variability around TARGET_WINS is explained by the model and a value of 100% indicates the model explains all of the variability related to the response variable.

Adjusted R-squared value

The Adjusted R-squared value for model 3 is 0.41. The Adjust R-squared value differs from R-squared in that it adjusts based on the number of predictors of TARGET_WINS in the model. However, the R-squared valued increases as the number of predictors of TOTAL_WINS increases.

The Adjust R-Squared value helps with determining whether including less predictors of TOTAL_WINS improves the model. A review on the Adjusted R-Squared values of models 1.1 and 2.1, which included more predictors than model 3.1, yields the same Adjusted R-Squared values.

Thus, the third model is still the best of the 3 models based on AIC and the Adjusted R-Squared values.

F-statistic

The F-statistic generated by model 1.3 is 106. The F-statistic compares the linear relationship between TOTAL_WINS and the predictor variable of the 3 models.

The higher F-statistic indicates a better fit of the linear relationship. A review of models 1.1 and 1.2 indicates a lower F-statistic 98.4 for model 1 and the same value of F-statistic of 106 for models 1.2 and 1.3.

Analysis of the AIC, R-squared, Adjusted R-squared and F-statistic indicates model 1.3 is the best model.

Standard Error

It is desirable that the standard error of each predictor be close to zero. A review of the standard error of the predictors of model 1.3 shows two predictors above 1 (PITCHING_SO_BB=3.03409, BATTING_BB_SO=3.47160). Since these 2 predictors are above 1 they will be removed and a new model will be developed to determine whether the removal improves the model.

Model 1.4

The predictors are: PITCHING_H + PITCHING_HR + PITCHING_BB + BATTING_3B + BATTING_HR + BATTING_BB + BATTING_SO + BATTING_TB + FIELDING_E + FIELDING_DP + BASERUN_SB

Removing predictors (PITCHING_SO_BB and BATTING_BB_SO) based on the Standard-error value greater than 1 did not improve the model. The AIC of model 1.4 is higher than model 1.3 and the R-squared and Adjusted R-squared values are lower.

Thus, model 1.3 is the best model to predict total number of wins based on backward regression approach.

We will now evaluate the validity of the model by analyzing the residuals.

Model Diagnostics

Visualizations of the residual values are used to determine whether, model 1.3, adheres to a linear relationship between response variable (TARGET_WINS) and the predictors. Residual values are the differences between the actual baseball statistics and the average of the baseball statics.

Pearson Residual Plots

These plots show whether there is a linear relationship and the strength of the relationship for each of the predictor variables. Since the R-squared value of the model is 0.41, it accounts for 41%

of the values that are around the line in the plot. Since there is no systematic pattern the model does have a linear relationship.

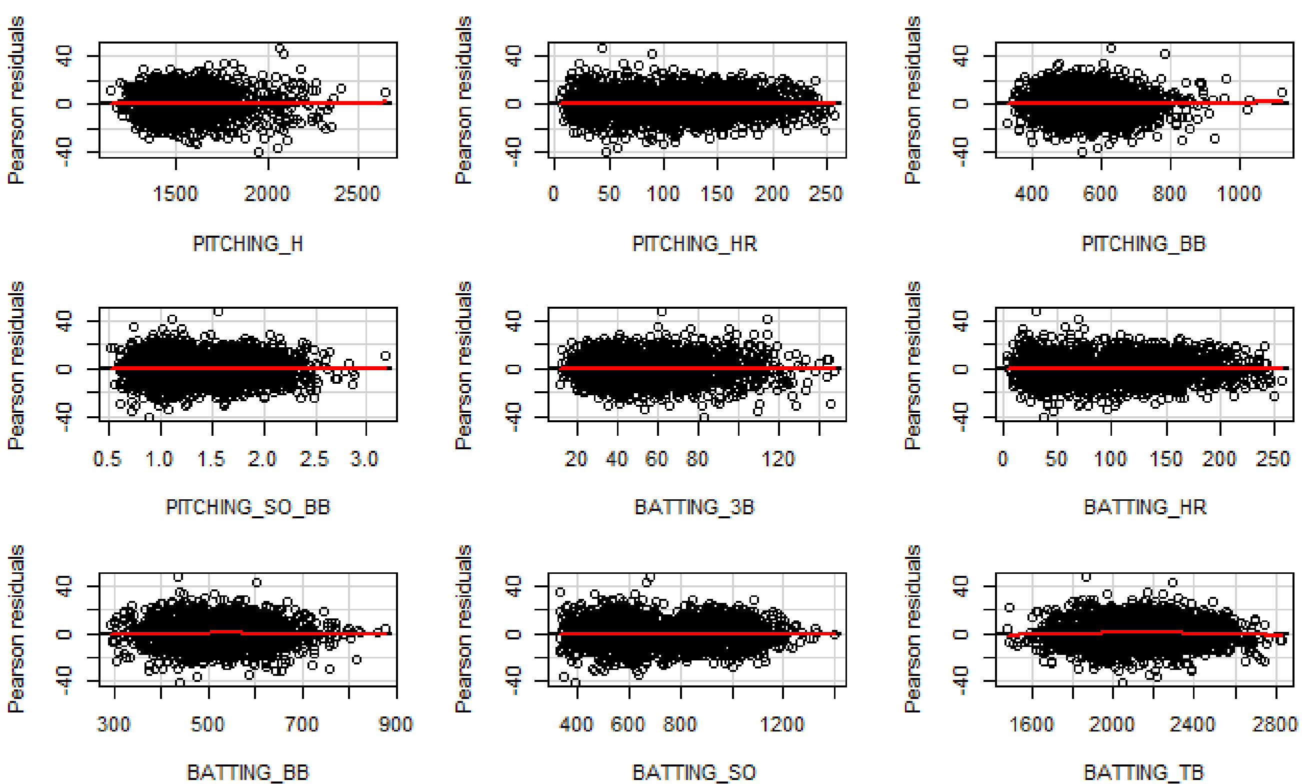
Standardized Residuals

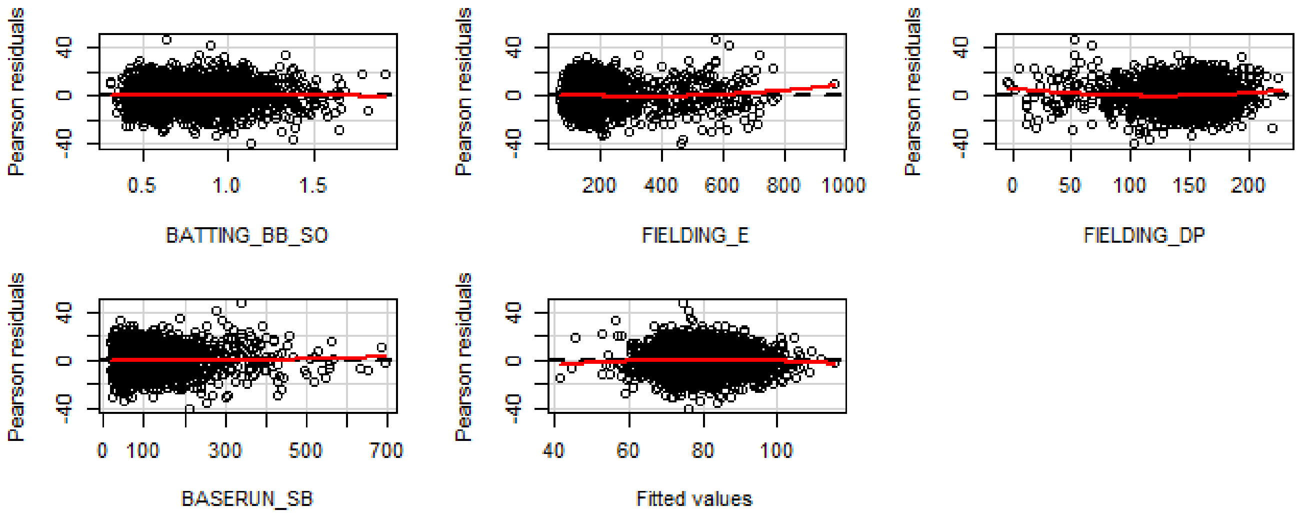
The standardized residual plot visualizes whether the data follow a normal distribution. A normal distribution shows whether the data is symmetric, bell shaped. Since the points are fall along the straight line the data are symmetric and bell shaped.

Cook's Distance Plot

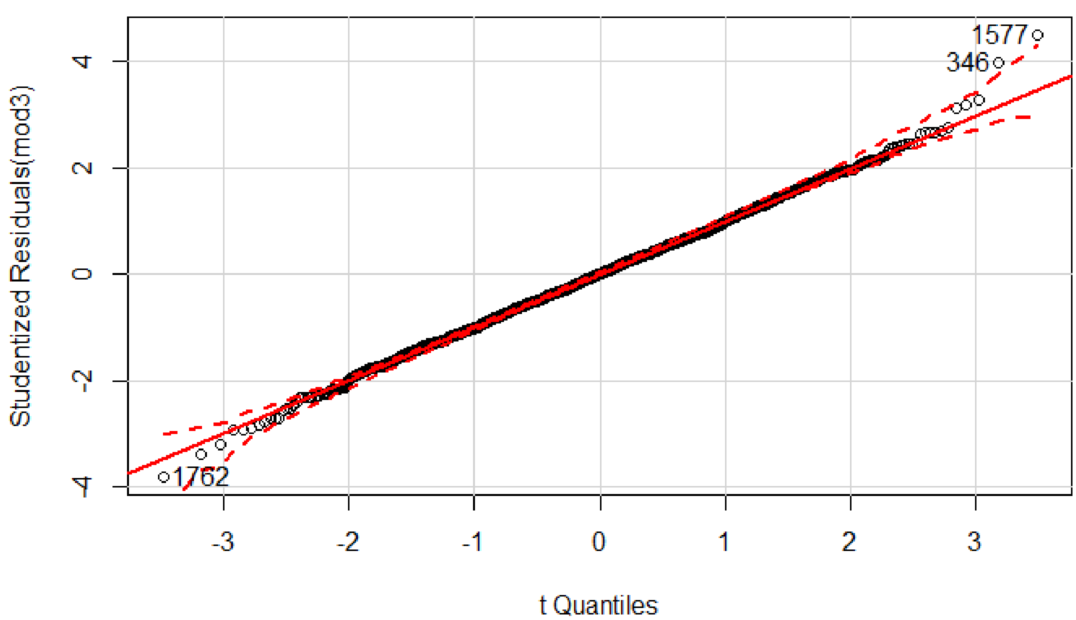
Cook's distance plot examines whether the distances of individual observations are considered influential to the quality of the model. The visualization suggests observations 1,377, 1,577 and 54 could be influential to the mode. However, the decision was made to retain these observations in the model.

Pearson Residual Plots for model 1.3:

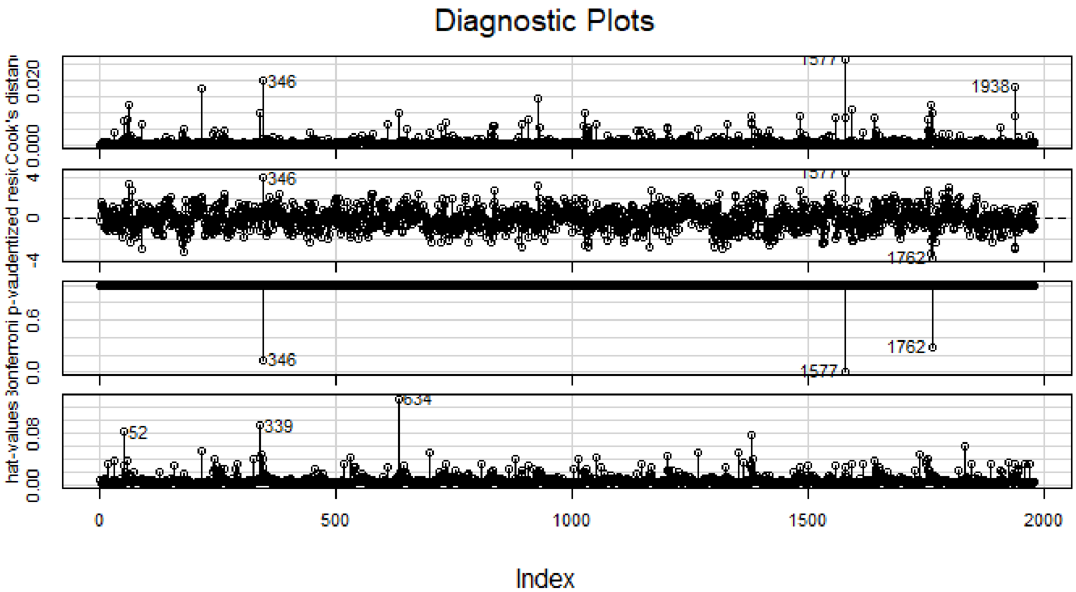


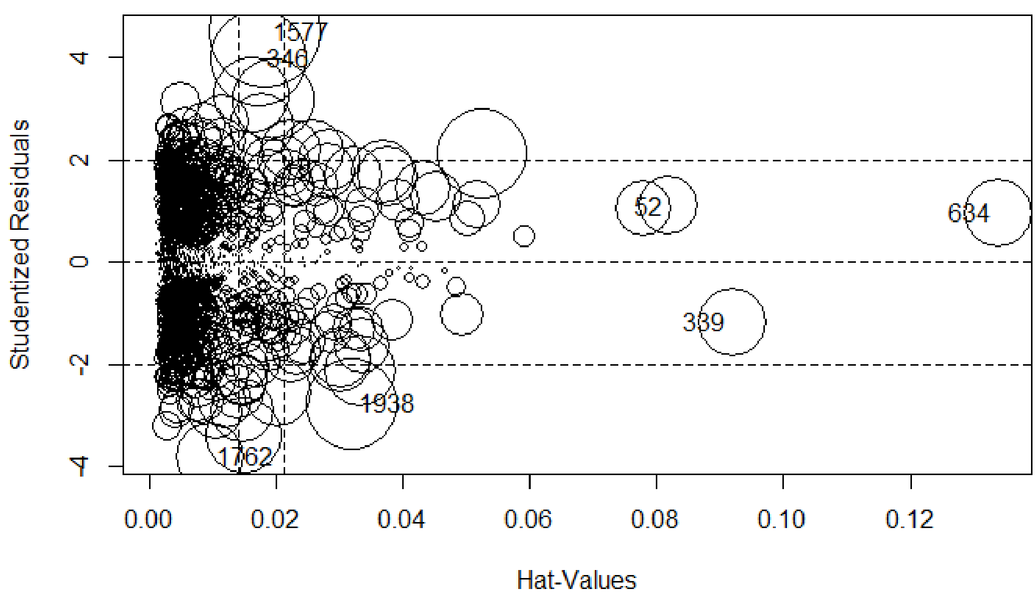


QQplot for model 1.3:

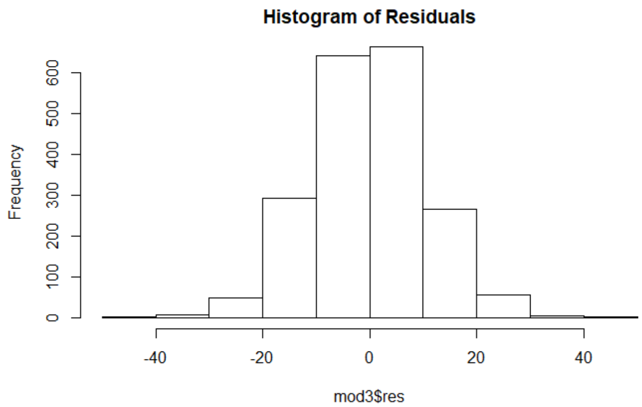


Cook's Distance Plot for model 1.3:





Histogram for Residuals for model 1.3



Model 1 Conclusion

Analysis of the residuals show that this model is a valid model for predicting total number of wins (TARGET_WINS)

VARIABLE	COEFFICIENT
Intercept	35.659404
PITCHING_H	0.072522
PITCHING_HR	-0.192645
PITCHING_BB	-0.152450
PITCHING_SO_BB	17.462216
BATTING_3B	0.235903
BATTING_HR	0.413415
BATTING_BB	0.267951
BATTING_SO	-0.069852
BATTING_TB	0.041007
BATTING_BB_SO	-18.168351
FIELDING_E	-0.101961
FIELDING_DP	-0.130086
BASERUN_SB	0.08801

The most significant variables in this model are PITCHING_SO_BB and BATTING_BB_SO, higher than most other offensive or defensive variables.

Model 2 - Total Base Model with forward selection

Using the variable Total bases as a starting point and adding additional variables until best model is reached.

This variable is calculated as follows:

TOTAL_BASE = BATTING_1B + 2xBATTING_2B + 3xBATTING_3B + 4xBATTING_HR

and is denoted as BATTING_TB in the data set.

Variable Selection

Apply Forward Stepwise Selection using BATTING_TB as the starting predictor variable. For this model, the base variables (those provided in the dataset) plus BATTING_TB and BATTING_1B will be considered.

Of note is that leaving the Batting statistics in the model (BATTING_1B, BATTING_2B, BATTING_3B, & BATTING_HR) yields a better model based on Adjusted R-squared and AIC values, despite likely collinearity among these variables with BATTING_TB.

Model with the selected variables based on the lowest AIC value

The resulting model includes the following ten predictor variables:

1. BATTING_TB
3. BASERUN_SB
4. FIELDING_E
5. BATTING_SO
6. FIELDING_DP
7. BATTING_BB
8. BATTING_2B
9. BATTING_3B
10. PITCHING_SO
11. PITCHING_BB
12. PITCHING_H
13. BATTING_1B
14. PITCHING_HR

Model Summary

```
Residuals:
    Min      1Q  Median      3Q     Max 
-42.118  -7.454  -0.004   7.162  45.808 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 50.737643   5.851964   8.670  < 2e-16 ***
BATTING_TB   0.093252   0.020723   4.500 7.20e-06 ***
BASERUN_SB   0.069571   0.004834  14.393  < 2e-16 ***
FIELDING_E  -0.099422   0.004654 -21.365  < 2e-16 ***
BATTING_SO  -0.081282   0.017310  -4.696 2.84e-06 ***
FIELDING_DP -0.125931   0.012554 -10.031  < 2e-16 ***
BATTING_BB   0.219999   0.028156   7.814 8.97e-15 ***
BATTING_2B  -0.260929   0.040790  -6.397 1.98e-10 ***
BATTING_3B  -0.153359   0.060471  -2.536 0.011288 *  
PITCHING_SO  0.059759   0.016398   3.644 0.000275 ***
PITCHING_BB -0.174829   0.025938  -6.740 2.07e-11 ***
PITCHING_H   0.056539   0.009994   5.657 1.77e-08 ***
BATTING_1B  -0.120548   0.020654  -5.837 6.22e-09 ***
PITCHING_HR -0.292367   0.076731  -3.810 0.000143 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.79 on 1965 degrees of freedom
Multiple R-squared:  0.4002,	Adjusted R-squared:  0.3963 
F-statistic: 100.9 on 13 and 1965 DF,  p-value: < 2.2e-16
```

BATTING_TB-based Regression Equation

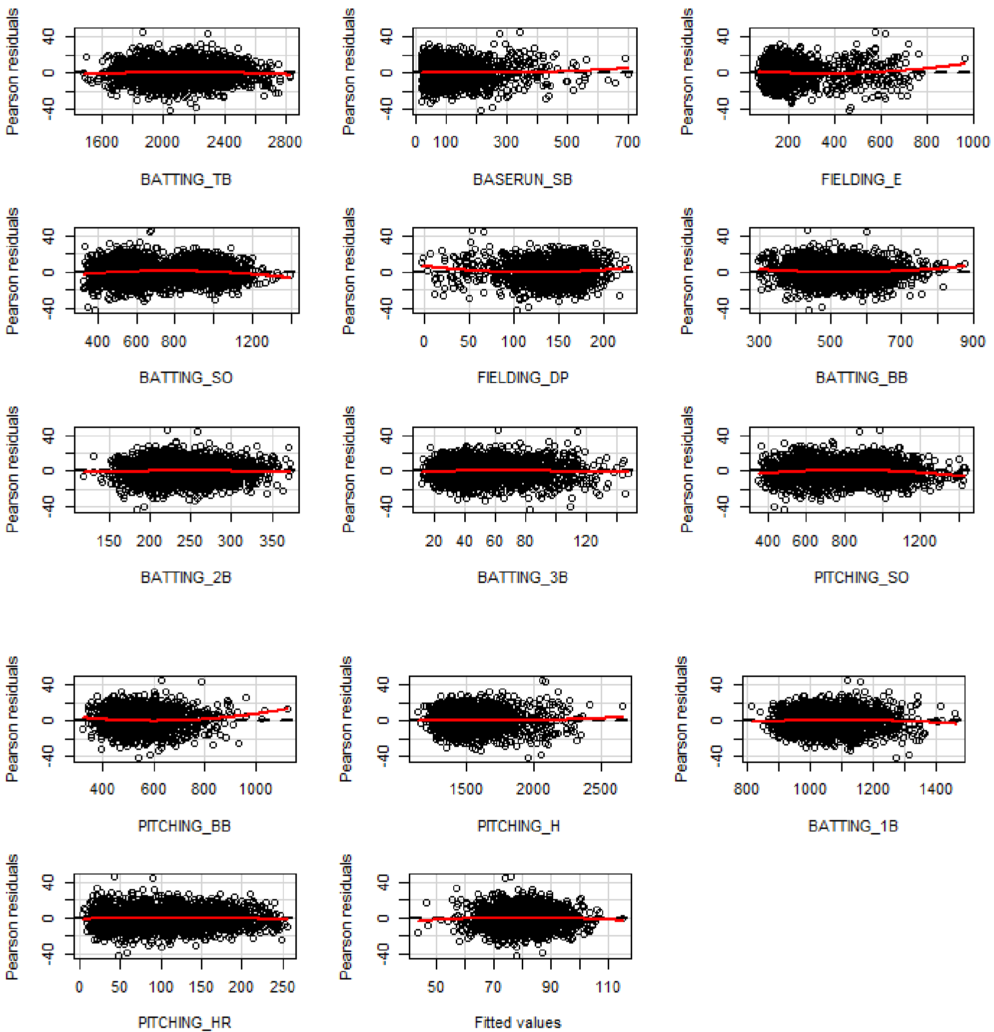
$$\begin{aligned}\widehat{\text{TARGET\_WINS}} &\\ = 50.74 + 0.09 \cdot \text{BATTING\_TB} + 0.70 \cdot \text{BASERUN\_SB} - 0.10 \cdot \text{FIELDING\_E} - 0.81 \cdot \text{BATTING\_SO} - 0.13.\,\text{FIELDING\_DP} - 0.11 \cdot \text{BATTING\_BB} - 0.26 \cdot \text{BATTING\_2B} - 0.15 \cdot \text{BATTING\_3B} &\\ + 0.06.\,\text{PITCHING\_SO} - 0.17.\,\text{PITCHING\_BB} + 0.06.\,\text{PTICHING\_H} - 0.12.\,\text{BATTING\_1B} - 0.29.\,\text{PITCHING\_HR} &\end{aligned}$$

Model Diagnostics

Examination of the residuals plot shows some indication of constant variance; however, the Residuals vs. Fitted plot may show a slight fanning appearance when looking from left to right.

Pearson Residual Plots for model 2:

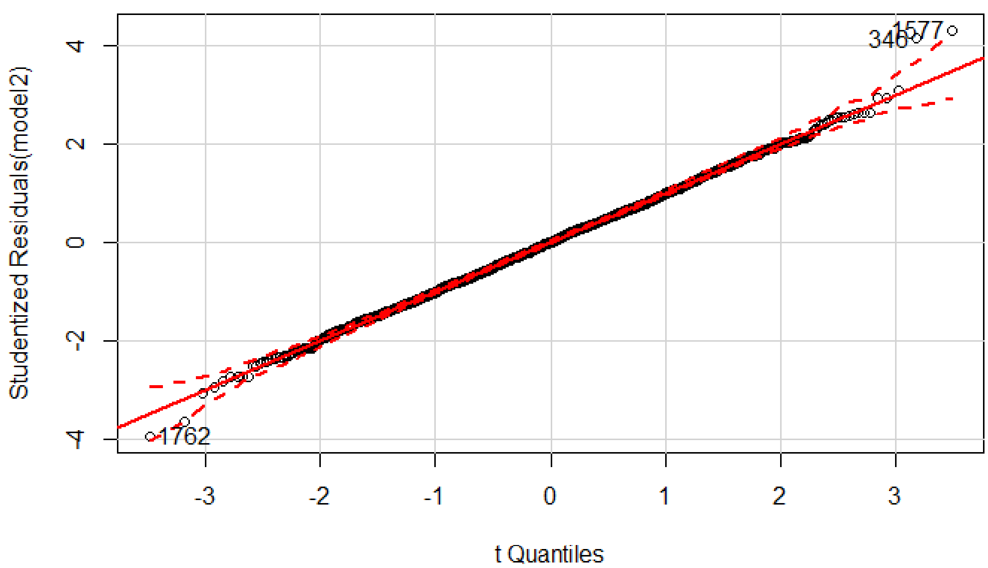
The residual plots seem to show constant variance.



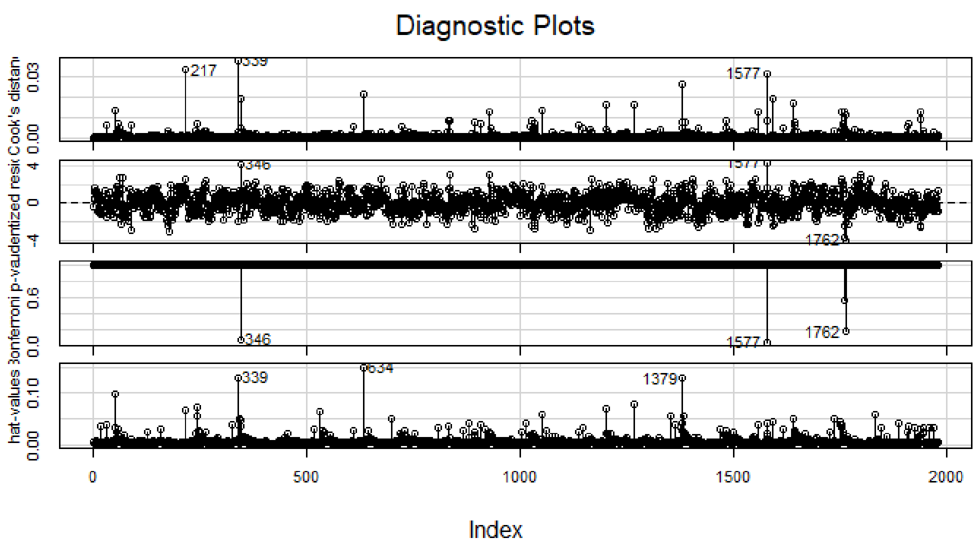
Additionally, the QQ plot looks mostly normal. There are three outliers labeled in the plot that may be contributing to fanning at the tails.

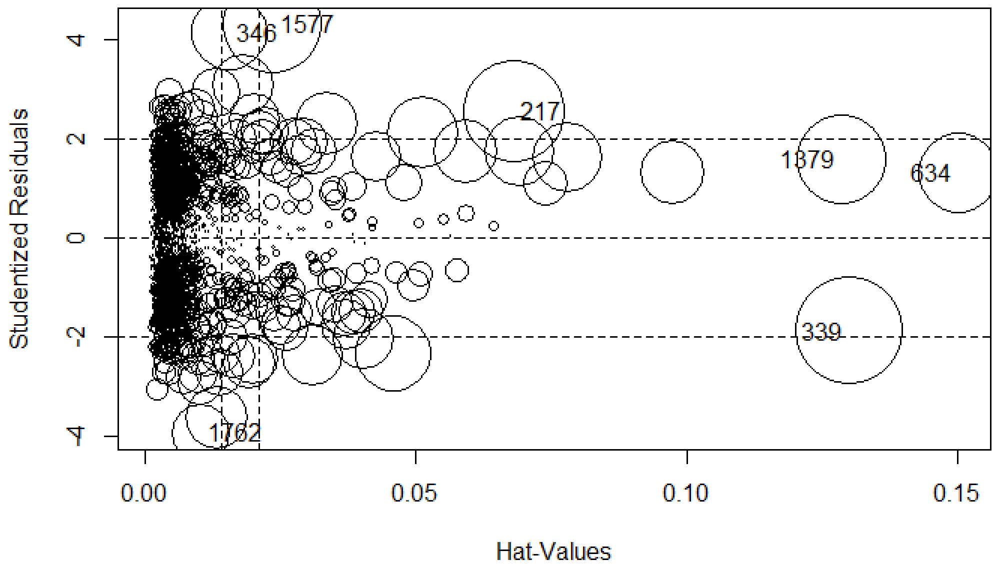
Outliers identified were at observations 1762, 346, and 1577. When looking further into these outliers there doesn't seem to be a huge indication of influence, but perhaps they can be removed to improve the model.

QQplot for model 2:

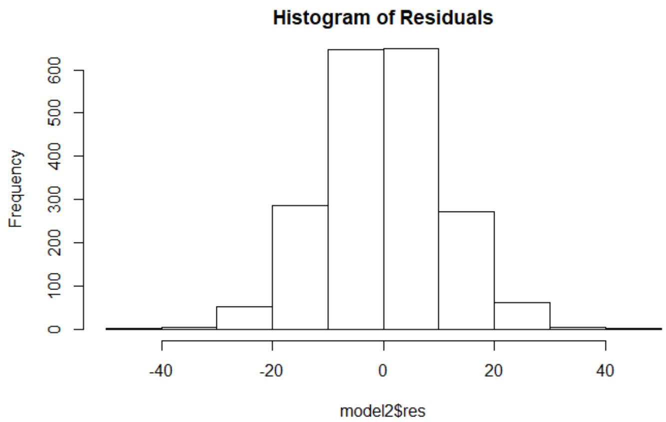


Cook's Distance Plot for model 2:





Histogram for Residuals for model 2:



Model 2 Conclusion

After reviewing the model, BATTING_BB (Walks) and BATTING_TB (Total Bases) were the two variables that played the most significant factors in wins. It is interesting to note that Walks was a larger contributor to wins in this model than Total Bases, which intuitively is a much more productive measure of scoring runs than walks alone.

VARIABLE	COEFFICIENT
Intercept	50.737643
BATTING_TB	0.093252
BASERUN_SB	0.069571
FIELDING_E	-0.099422
BATTING_SO	-0.081282
FIELDING_DP	-0.125931
BATTING_BB	0.219999
BATTING_2B	-0.260929
BATTING_3B	-0.153359
PITCHING_SO	0.059759
PITCHING_BB	-0.174829
PITCHING_H	0,056539
BATTING_1B	-0.120548
PITCHING_HR	-0.292367

Model 3 - Walks and Hits Per Game Played (WHGP)

WHIP (Walks plus Hits per Inning Pitched) is a saber metric measure of how many runners a pitcher has allowed per inning pitched. Since the innings pitched statistic is not provided with the given dataset, the WHIP statistic has been modified to be "Walks plus Hits per Game Played." `WHGP` is a statistic is a measure of a team's success in preventing a batter from reaching base. From a defensive perspective, a lower WHGP score indicates better performance in preventing batters from reaching base whereas, from an offensive perspective, a higher score indicates a propensity to let batters on base.

Model 3 will focus on WHGP as a predictor by examining the significance of this metric specifically in combination with other offensive and defensive metrics to determine the optimal regression model for predicting wins.

$$WHGP = (PITCHING\_H + PITCHING\_BB)/162$$

where:
PITCHING_H is the number of hits a team allowed and
PITCHING_BB is the number of walks a team allowed

The number of games played is set to 162

Variable Selection

Apply Forward Stepwise Selection using WHGP as the starting predictor variable. For this model, the base variables (those provided in the dataset) plus BATTING_TB and BATTING_1B will be considered. Due to the collinearity between the BATTING-related predictor variables, BATTING_TB will be used in place of BATTING_1B, BATTING_2B, BATTING_3B, and BATTING_HR.

Of note is that leaving PITCHING_BB and PITCHING_H in the model yields a better model based on Adjusted R-squared and AIC values, despite likely collinearity among these variables with WHGP.

Model with the selected variables based on the lowest AIC value

The resulting model includes the following ten predictor variables:

1. WHGP
2. BATTING_TB
3. BASERUN_SB
4. FIELDING_E
5. PITCHING_SO
6. FIELDING_DP
7. BATTING_BB
8. BATTING_SO
9. PITCHING_BB
10. PITCHING_HR

Model Summary

Reviewing the model summary, BATTING_TB is not proving to be a significant predictor variable. Consequently, this variable will be dropped from the regression model. After updating the model and re-examining, the AIC value drops slightly and adjusted R-squared increases slightly.

```
Residuals:
    Min      1Q  Median      3Q     Max 
-41.068  -7.703   0.046   7.340  46.801 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 53.509465   5.343901  10.013  < 2e-16 ***
WHGP         4.812634   0.425928  11.299  < 2e-16 ***
BASERUN_SB   0.073945   0.004943  14.961  < 2e-16 ***
FIELDING_E  -0.077672   0.004327 -17.950  < 2e-16 ***
PITCHING_SO  0.035557   0.012467   2.852  0.00439 ** 
FIELDING_DP -0.134489   0.012843 -10.472  < 2e-16 ***
BATTING_BB   0.179262   0.021777   8.232 3.32e-16 ***
BATTING_SO  -0.060580   0.013318  -4.549 5.73e-06 ***
PITCHING_BB -0.169435   0.020887  -8.112 8.66e-16 ***
PITCHING_HR  0.070050   0.008447   8.293  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.15 on 1969 degrees of freedom
Multiple R-squared:  0.3586,	Adjusted R-squared:  0.3557 
F-statistic: 122.3 on 9 and 1969 DF,  p-value: < 2.2e-16
```

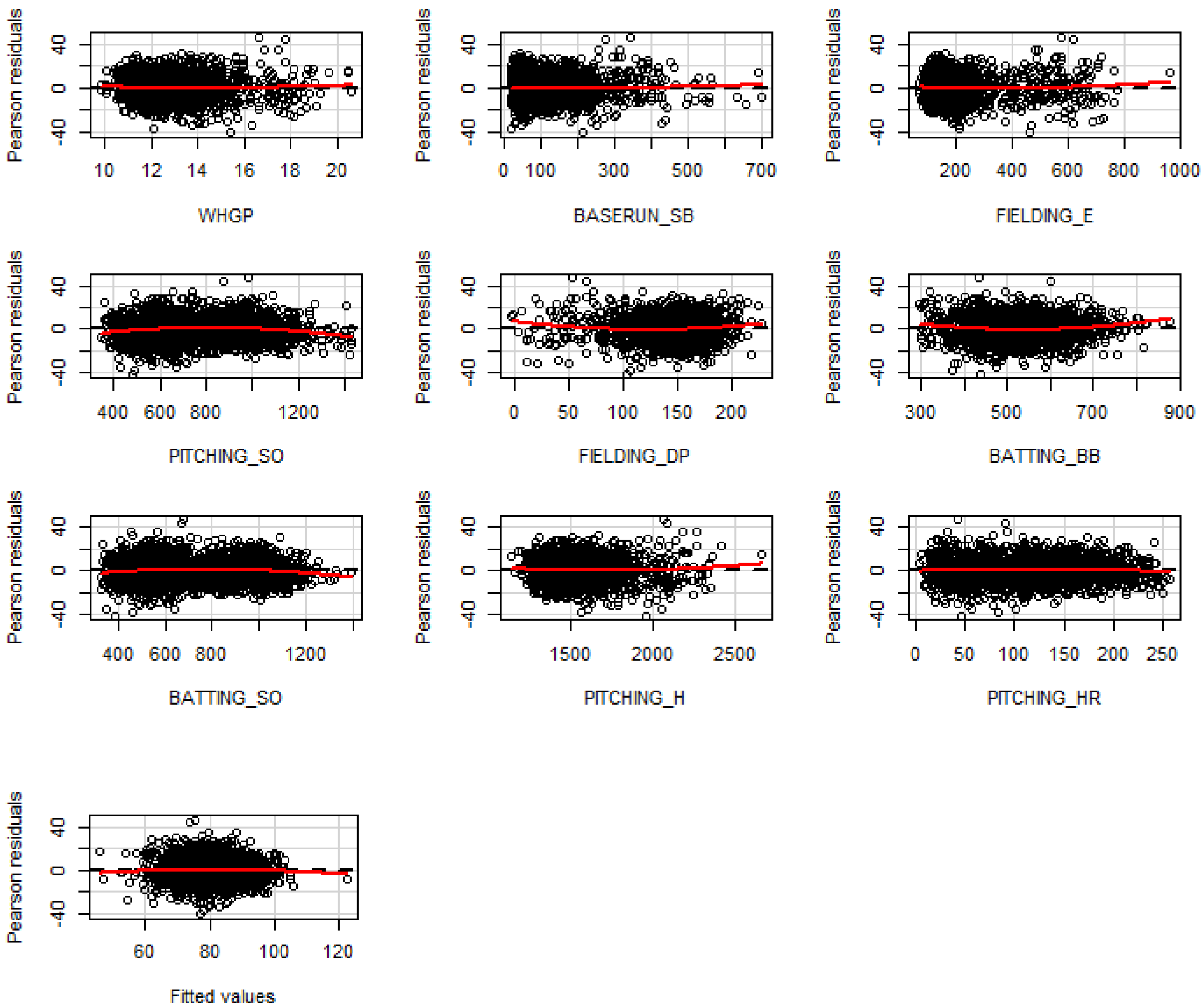
WHGP-based Regression Equation

$$\widehat{\mathrm{TARGET\_WINS}} = 53.50 + 4.81 \cdot \mathrm{WHGP} + 0.073 \cdot \mathrm{FIELDING\_E} + 0.035 \cdot \mathrm{PITCHING\_SO} + 0.035 \cdot \mathrm{FIELDING\_DP} + 0.18 \cdot \mathrm{BATTING\_BB} - 0.06 \cdot \mathrm{BATTING\_SO} - 0.17 \cdot \mathrm{PITCHING\_BB} + 0.07 \cdot \mathrm{PITCHING\_HR}$$

Model Diagnostics

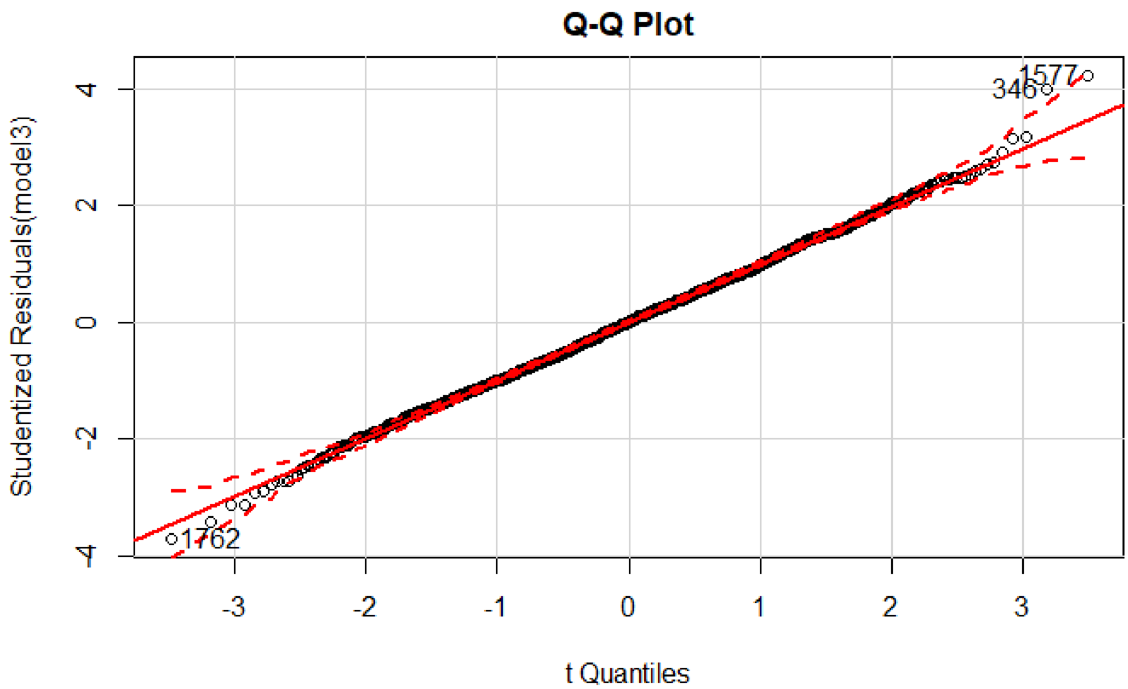
Examination of the residuals plot shows some indication of constant variance; however, the Residuals vs. Fitted plot may show a slight fanning appearance when looking from left to right.

Pearson Residual Plots for model 3:



However, the Q-Q Plot of the standardized residuals looks very close to normal with a few noted outliers in the tails -- observations 1577, 346, and 1762. These observations may be contributing to the slight fanning in the Residual vs. Fitted Plot.

QQplot for model 3:



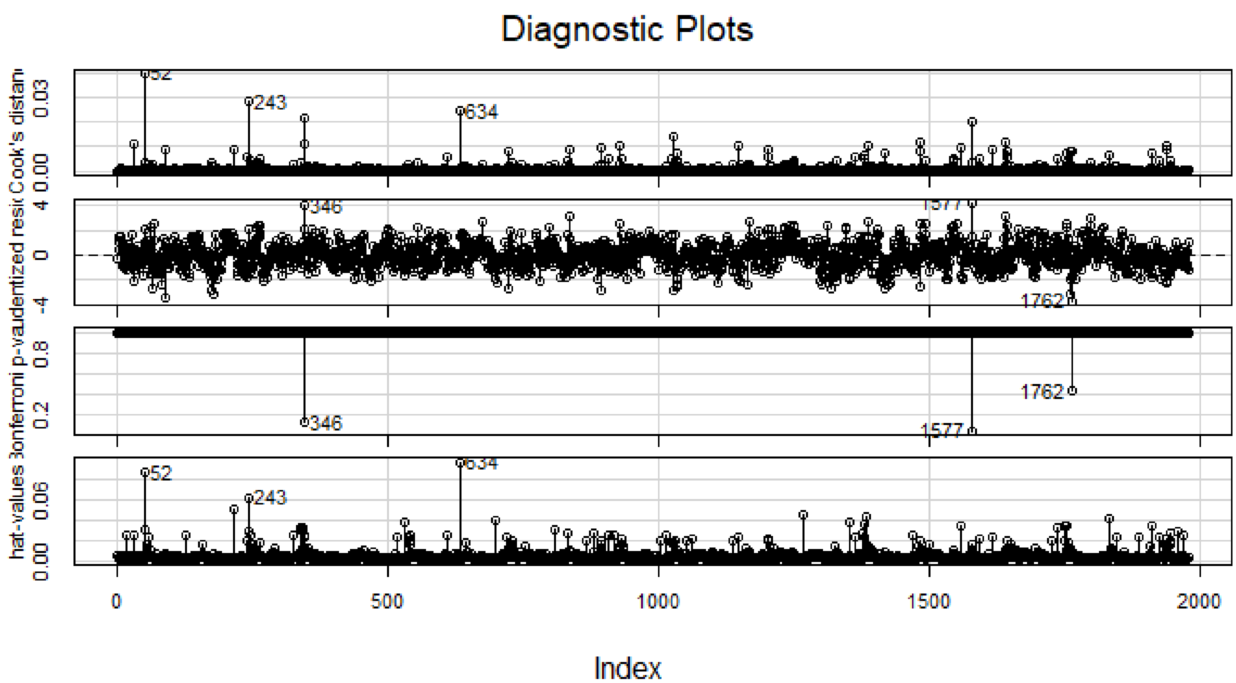
Transformations

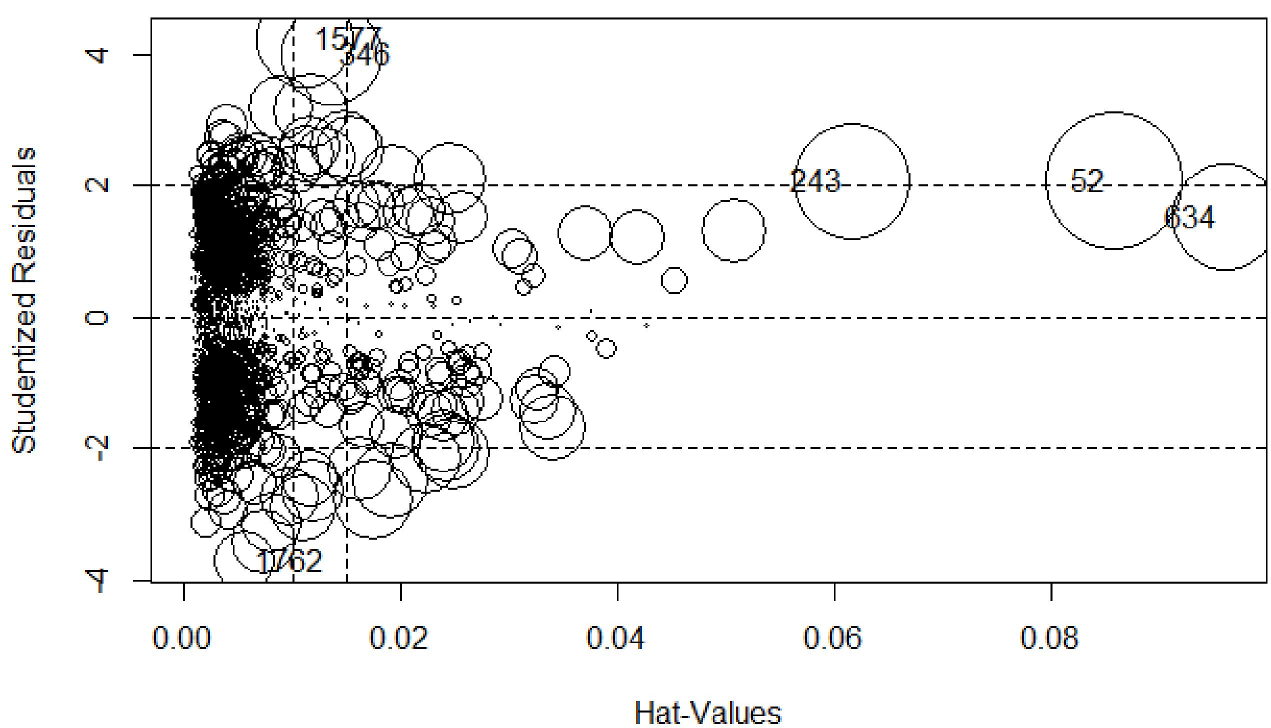
Several options for transformations such as a power transformation or reciprocal values of predictors (FIELDING_E and FIELDING_DP) were applied. However, none yielded better models as determined by the resulting Adjusted R-squared and F-statistics.

Outliers & Influence Points

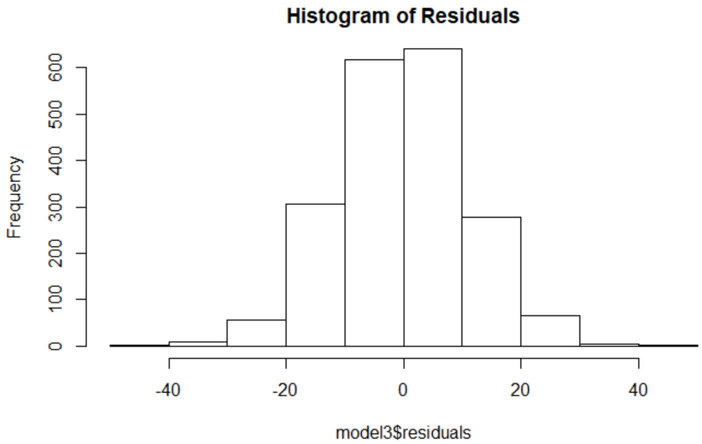
Looking for the overlap of outliers and high leverage points, we see that three observations in particular (observations 346, 1577, and 1762) may be impacting the model. These observations may be candidates for removal.

Cook's Distance Plot for model 3:





Histogram for Residuals for model 3:



Model 3 Conclusion

Among the predictor variables included in Model 3, Walks and Hits Per Game Played is the most significant contributor to wins with a coefficient of nearly 5. From an offensive perspective, a team playing against a team with a high WHGP metric will be more likely to win. However, WHGP is based on pitching statistics so it is counter-intuitive that a higher WHGP would have a positive relationship with winning. Intuitively a team that allows more of its opponent's batters on base (reflected in a high WGHP) would be more likely to lose.

VARIABLE	COEFFICIENT
Intercept	53.50946
WHGP	4.812634
BASERUN_SB	0.073945
FIELDING_E	-0.077672
PITCHING_SO	0.035557
FIELDING_DP	-0.134489
BATTING_BB	0.179262
BATTING_SO	-0.060580
PITCHING_BB	-0.169435
PITCHING_HR	0.070050

Model 4 - BSR Model (SaberMetrics Model)

Model 4 – BSR forward

In this model we incorporate our calculated metric Base Runs (BsR), a sabermetric stat created by David Smyth, to predict the number of runs a team would be expected to have scored based on the types of hits and number of walks.

BsR is calculated as follows:

$$BsR = \frac{A \cdot B}{B + C} + D$$

$$A = Hits + Walks - Homeruns$$

$$B = 1.4 \times Total\ Bases - 0.6 \times Hits - 3 \times Homeruns + 0.1 \times Walks$$

$$\mathrm{C} = \mathrm{A} \cdot \mathrm{B} - \mathrm{Hits}$$

$$D = Homeruns$$

Variable Selection

We include all available variables, beginning with BsR, and use forward stepwise regression to add statistically significant variables to the model.

Model with the selected variables based on the lowest AIC value

The resulting model includes the following ten predictor variables:

1. BsR
2. BASERUN_SB
3. FIELDING_E
4. BATTING_SO
5. FIELDING_DP
6. BATTING_2B
7. BATTING_1B
8. BATTING_3B
9. PITCHING_H
10. BATTING_BB_SO
11. PITCHING_HR
12. PITCHING_SO_BB
13. BATTING_BB
14. WHGP
15. BATTING_TB

Model Summary

As variables were added to the model, the statistical significance of some initial variables was reduced. In fact, our main statistic of interest, BsR, is no longer statistically significant. To develop the best selection of variables, we are also incorporating a bidirectional method to revisit the significance of variables added earlier in the analysis.

```
Residuals:
    Min      1Q  Median      3Q     Max 
-40.364  -6.939   0.007   6.936  47.465 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    34.156250  36.427508   0.938  0.34854    
BsR            -0.001671   0.061891  -0.027  0.97847    
BASERUN_SB      0.068738   0.004770  14.412  < 2e-16 ***
FIELDING_E     -0.101917   0.004641 -21.961  < 2e-16 ***
BATTING_SO     -0.069798   0.007487  -9.323  < 2e-16 ***
FIELDING_DP    -0.130358   0.012455 -10.467  < 2e-16 ***
BATTING_2B     -0.080131   0.056021  -1.430  0.15276    
BATTING_1B     -0.038194   0.037585  -1.016  0.30966    
BATTING_3B      0.116019   0.076728   1.512  0.13068    
PITCHING_H      0.071390   0.008683   8.221 3.61e-16 ***
BATTING_BB_SO -18.207216   3.635818  -5.008 6.00e-07 ***
PITCHING_HR    -0.191548   0.065185  -2.939  0.00334 ** 
PITCHING_SO_BB 17.497521   3.116173   5.615 2.24e-08 ***
BATTING_BB      0.265593   0.037325   7.116 1.56e-12 ***
PITCHING_BB    -0.149573   0.025982  -5.757 9.92e-09 ***
BATTING_HR      0.252325   0.112052   2.252  0.02444 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.69 on 1963 degrees of freedom
Multiple R-squared:  0.4122,	Adjusted R-squared:  0.4077 
F-statistic: 91.77 on 15 and 1963 DF,  p-value: < 2.2e-16
```

Model 4B – BsR Bidirectional

Variable Selection

We include all available variables, beginning with BsR, and use bi-directional stepwise regression to revisit the significance of variables added earlier in the analysis.

Model with the selected variables based on the lowest AIC value

The resulting model includes the following ten predictor variables:

1. BASERUN_SB
2. FIELDING_E
3. BATTING_SO
4. FIELDING_DP
5. BATTING_2B
6. BATTING_1B
7. PITCHING_H
8. BATTING_BB_SO
9. PITCHING_HR
10. PITCHING_SO_BB
11. BATTING_BB
12. WHGP
13. BATTING_TB

Model Summary

Using the bidirectional approach, BsR was removed from the model. Once BsR was removed, BATTING_2B, BATTING_1B and BATTING_3B regained their significance. This is likely caused by collinearity within the variables as BsR is a derived stat based on large part on hits. Because BsR was found to not add predictive ability to our model, Model 4B is the superior model with a higher F-statistic and slightly improved adjusted R squared and AIC values.

```
Residuals:
    Min      1Q  Median      3Q     Max 
-40.357  -6.938   0.012   6.944  47.467 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     35.120603   7.137530   4.921 9.35e-07 ***
BASERUN_SB       0.068748   0.004754  14.460  < 2e-16 ***
FIELDING_E      -0.101922   0.004634 -21.993  < 2e-16 ***
BATTING_SO      -0.069789   0.007478  -9.333  < 2e-16 ***
FIELDING_DP     -0.130336   0.012425 -10.490  < 2e-16 ***
BATTING_2B      -0.081610   0.011716  -6.966 4.43e-12 ***
BATTING_1B      -0.039170   0.010263  -3.817 0.000139 ***
BATTING_3B       0.114012   0.018959   6.014 2.16e-09 ***
PITCHING_H       0.071381   0.008675   8.228 3.41e-16 ***
BATTING_BB_SO  -18.235017   3.486053  -5.231 1.87e-07 ***
PITCHING_HR     -0.191683   0.064976  -2.950 0.003215 ** 
PITCHING_SO_BB  17.478624   3.035769   5.758 9.88e-09 ***
BATTING_BB       0.264966   0.029219   9.068  < 2e-16 ***
PITCHING_BB     -0.149522   0.025908  -5.771 9.12e-09 ***
BATTING_HR       0.249909   0.067422   3.707 0.000216 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.69 on 1964 degrees of freedom
Multiple R-squared:  0.4122,	Adjusted R-squared:  0.408 
F-statistic: 98.38 on 14 and 1964 DF,  p-value: < 2.2e-16
```

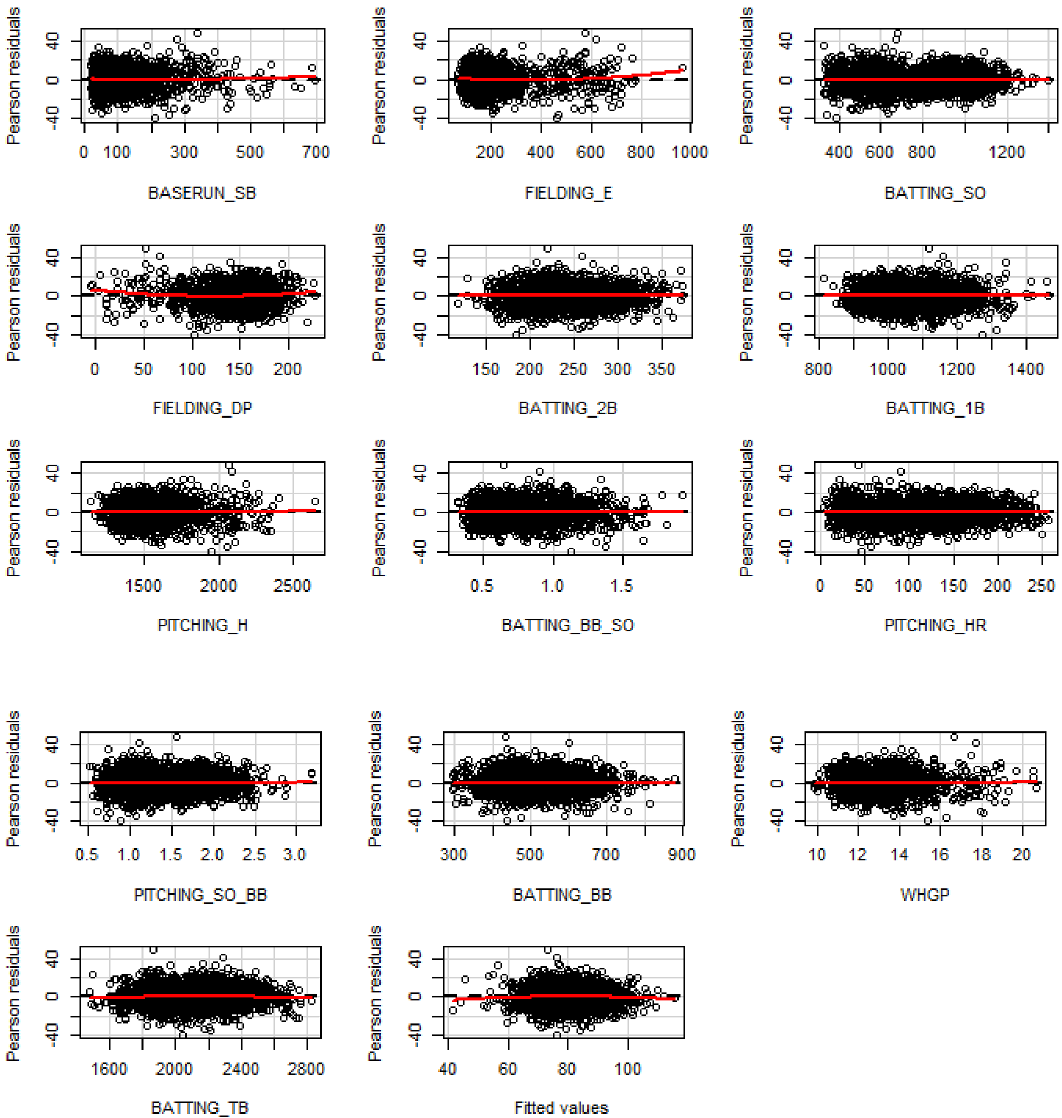
Model4b-based Regression Equation

$$\widehat{\mathrm{TARGET\_WINS}} = 34.34 + 0.07\mathrm{BASE\_SB} - 0.10\mathrm{FIELDING\_E} - 0.69\mathrm{BATTING\_SO} - 0.1\mathrm{FIELDING\_DP} - 0.08\mathrm{BATTING\_2B} - 0.04\mathrm{BATTING\_1B} + 0.07\mathrm{PITCHING\_H} - 18.24\mathrm{BATTING\_BB\_SO} - 0.19\mathrm{PITCHING\_HR} + 17.48\mathrm{PITCHING\_SO\_BB} + 0.26\mathrm{BATTING\_BB} - 0.15\mathrm{PITCHING\_BB} + 0.25\mathrm{BATTING\_HR}$$

Model Diagnostics

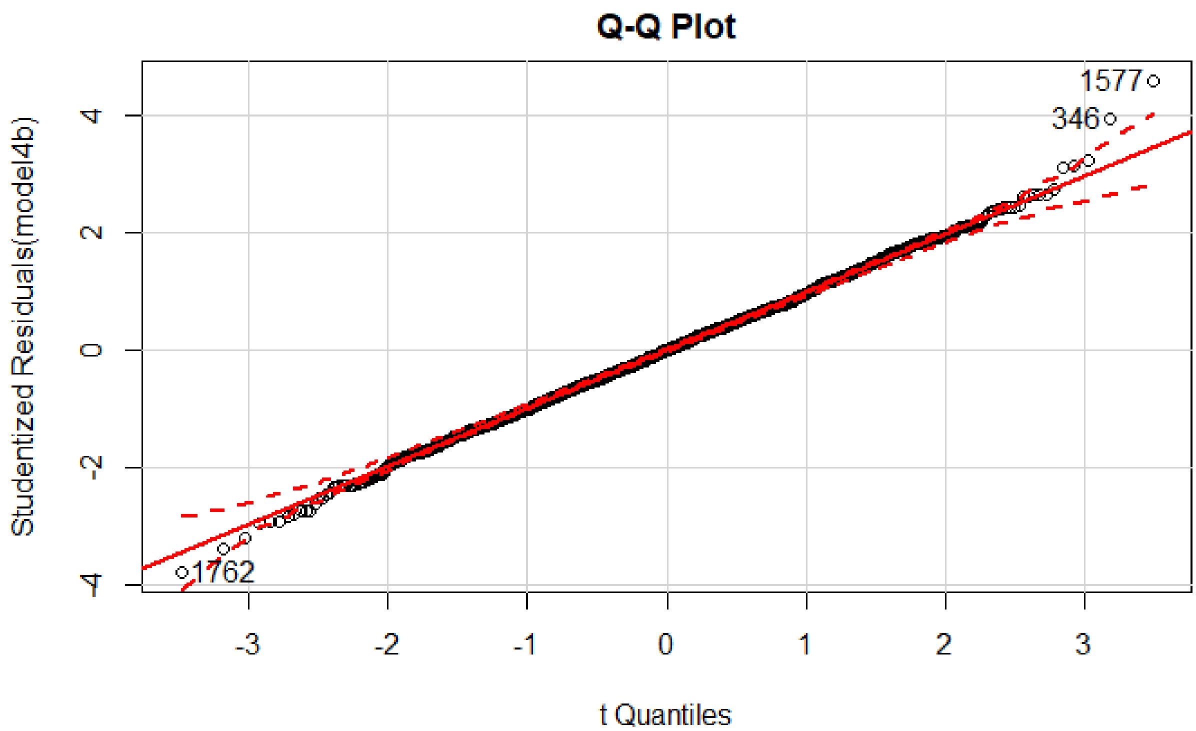
Examination of the residuals plot shows some indication of constant variance. However, the Residuals vs. Fitted plot may show a slight fanning appearance when looking from left to right.

*Pearson Residual Plots for model*4b

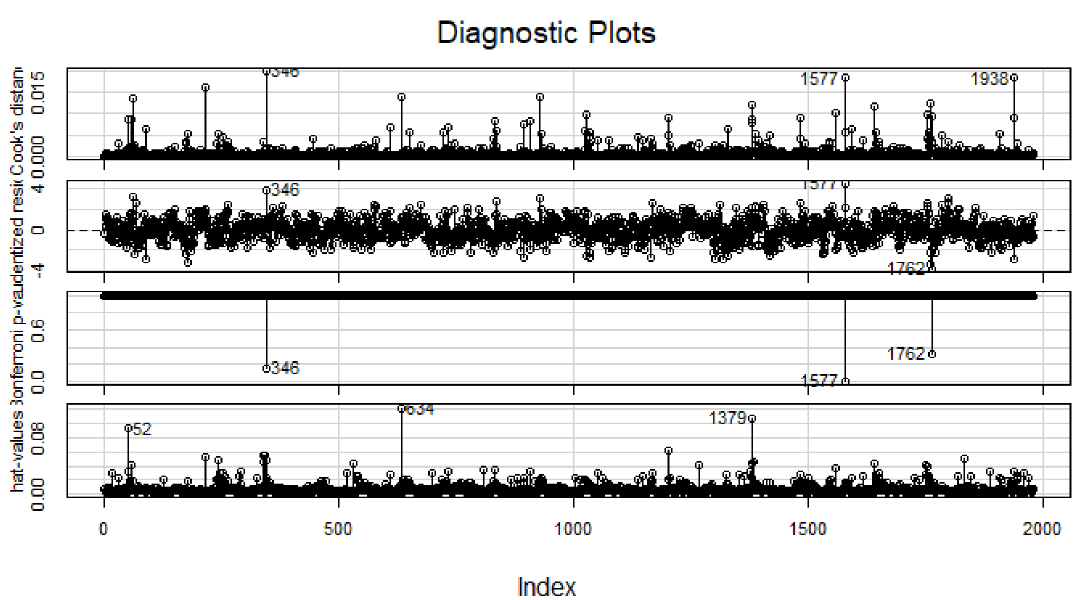


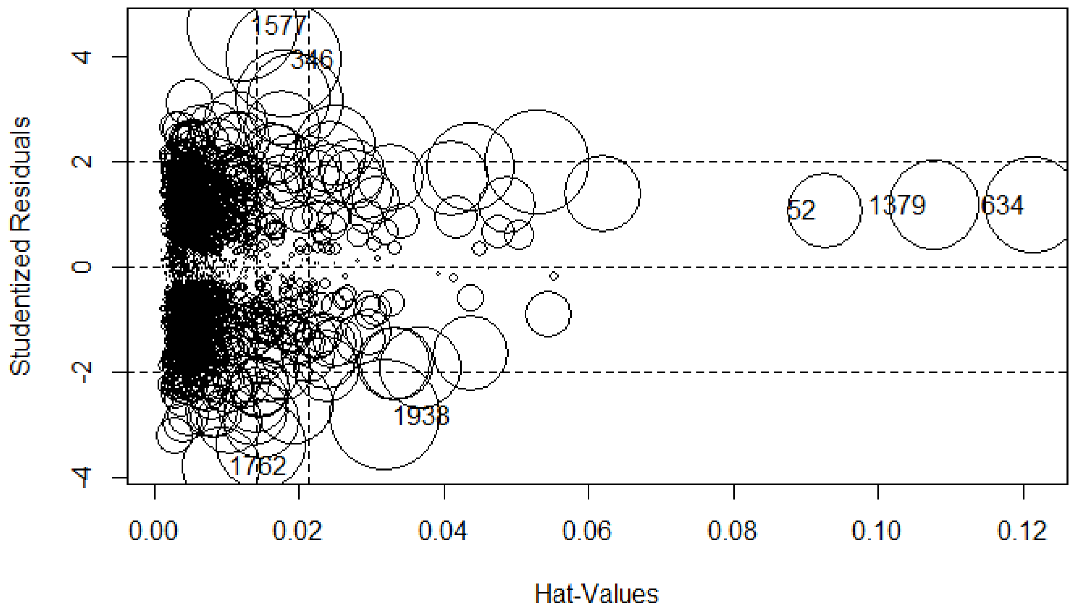
However, the Q-Q Plot of the standardized residuals looks very close to normal with a few noted outliers in the tails -- observations 1577, 346, and 1762. These observations may be contributing to the slight fanning in the Residual vs. Fitted Plot.

QQplot for model 4b

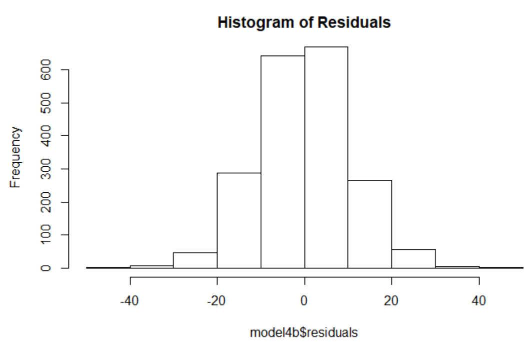


Cook's Distance Plot for model 4b:





Histogram for Residuals for model 4b:



Model 4b Conclusion

We have found that BsRis not significant in the bidirectional model.

VARIABLE	COEFFICIENT
Intercept	35.120603
BASERUN_SB	0.068748
FIELDING_E	-0.101922
BATTING_SO	0.069789
FIELDING_DP	-0.130336
BATTING_2B	-0.081610
BATTING_1B	-0.039170
BATTING_3B	0.114012
PITCHING_H	0.071381
BATTING_BB_SO	-18.235017
PITCHING_HR	-0.191683
PITCHING_SO_BB	17.478624
BATTING_BB	0.264966
PITCHING_BB	-0.149522
BATTING_HR	0.249909

Model Selection

The key statistics from the four models are summarized below.

Model	Residual Standard Error	Adjusted R-squared	F statistic	AIC	Predicted Accuracy (Train)
Model 1	10.69	0.4083	106	9390	90.1%
Model 2	10.79	0.3963	100.9	9430	90%
Model 3	11.15	0.3557	122.3	9555	89.6%
Model 4	10.69	0.4080	98.38	9392	90.1%

The statistics for all four models are quite close. On the strength of lower AIC value, higher Adjusted R-squared, and higher Predictive Accuracy, Model 1 is selected as the final model for predicting total wins.

Using our model to make prediction

We will now obtain the evaluation set and evaluate it briefly. We will also apply the same transformations to the evaluation set in terms of matching our final predictors and imputation of missing variables, our resulting evaluation data set will be loaded to Github for reproducibility of results.

https://raw.githubusercontent.com/vbriot28/Data621_group2/master/data_group2_evaluation_nbc.csv

The results of the prediction are reasonable as compare with training set:

Dataset	Min	1stQtr	Median	Mean	3rdQtr	Max
Train	0	71	82	80.79	92	146
Train (transformed)	27	72	82	80.92	91	123
Evaluation	36	74	81	80.24	87	119

The full prediction results can be found at:

https://github.com/vbriot28/Data621_group2/blob/master/data_group2_prediction_model1.csv

Areas for Further Study

There is a question as to whether our test set contains years that are more current while our training set contains all of the earlier data. If that is the case, the game of baseball was much different in the past, with less power-hitters just being one example. It would be helpful to perform further analysis using time periods to determine if data before the modern era is useful to predict current success. Furthermore, additional transformations could have reduced the skew of some of the variable distributions. In our early analysis, the transformations we performed did not improve model predictability; however, this could be investigated further.

Conclusion

The methodology applied in the project (EDA, Data Preparation, and Model Building) resulted in the creation of four multiple linear regression models for predicting baseball wins. The four modeling attempts purposively used slightly different approaches on the same set of transformed data to find an optimal model. All four models revealed and helped quantify significant characteristics of baseball statistics relevant to winning -- most were consistent with intuition while some were not, as in the case of the positive relationship between Wins and Walks plus Hits per Game. In the end, Model 1 was selected as the best candidate model based on having the lower AIC value, highest Adjusted R-squared, and highest accuracy score (tied).

The predicted values resulting from Model 1 are in line with wins in the training dataset based on a five-number summary comparison. Evaluating the prediction performance of the selected model against actual wins in the evaluation dataset would be a next step in further diagnosing and refining the model.

References

http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm
http://www.stat.columbia.edu/~gelman/arm/missing.pdf
https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/
https://www.r-bloggers.com/missing-value-treatment/https://datascienceplus.com/imputing-missing-data-with-r-mice-package/
https://www.r-bloggers.com/unsupervised-data-pre-processing-individual-predictors/
https://www.r-bloggers.com/computing-and-visualizing-pca-in-r/
http://www.baseball-almanac.com/rb_menu.shtml

https://sites.ualberta.ca/~lkgray/uploads/7/3/6/2/7362679/slides_-_multiplelinearregressionaic.pdf

https://www.youtube.com/watch?v=TzhgPXrFSm8

http://blog.minitab.com/blog/adventures-in-statistics-2/what-is-the-f-test-of-overall-significance-in-regression-analysis

http://rstatistics.net/robust-regression/

APPENDIX – R Code

1.) R Packages

```
library(psych)
library(ggplot2)
library(reshape2)
library(pastecs)
library(mice)
library(VIM)
library(corrplot)
library(dplyr)
library(DataExplorer)
library(caret)
library(MASS)
library(car)
```

2.) Read the dataset

```
data_train_moneyball <-
read.csv("https://raw.githubusercontent.com/vbriot28/Data621_group2/ma
ster/moneyball-training-data.csv", header = TRUE)

colnames(data_train_moneyball) = gsub("TEAM_", "",
colnames(data_train_moneyball))
```

3.) Data Exploration

```
Variable_names <- c("INDEX", "TARGET_WINS", "BATTING_H", "BATTING_2B",
"BATTING_3B", "BATTING_HR", "BATTING_BB", "BATTING_HBP", "BATTING_SO",
"BASERUN_SB", "BASERUN_CS", "TFIELDING_E", "FIELDING_DP",
"PITCHING_BB", "PITCHING_H", "PITCHING_HR", "PITCHING_SO")

Definitions <- c("Identification Variable", "Number of wins", "Base
Hits by batters (1B,2B,3B,HR)", "Doubles by batters (2B)", "Triples by
batters (3B)", "Homeruns by batters (4B)", "Walks by batters",
"Batters hit by pitch", "Batters hit by pitch", "Stolen bases",
"Caught stealing", "Errors", "Double Plays", "Walks allowed", "Hits
allowed", "Homeruns allowed", "Strikeouts by pitchers")

Theoritical_effect <- c("None", "", "Positive", "Positive",
"Positive", "Positive", "Positive", "Positive", "Negative",
"Positive", "Negative", "Negative", "Positive", "Negative",
"Negative", "Negative", "Positive")

Category <- c("Identifier", "Result", "Batting", "Batting", "Batting",
"Batting", "Batting", "Batting", "Batting", "Baserunning",
"Baserunning", "Fielding", "Fielding", "Pitching", "Pitching",
"Pitching", "Pitching")

Variable_type <- c("", "Response", "Predictor", "Predictor",
"Predictor", "Predictor", "Predictor", "Predictor", "Predictor",
"Predictor", "Predictor", "Predictor", "Predictor", "Predictor",
"Predictor", "Predictor", "Predictor")
```

```
Data_type <- c("", "Count", "Count", "Count", "Count", "Count",
"Count", "Count", "Count", "Count", "Count", "Count", "Count",
"Count", "Count", "Count", "Count")

df_moneyball_md <- cbind.data.frame (Variable_names, Definitions,
Theoritical_effect, Category, Variable_type, Data_type)

colnames(df_moneyball_md) <- c("Variable Name", "Definition",
"Theoritical Effect", "Category", "Variable Type", "Data Type")

knitr::kable(df_moneyball_md)
```

EDA

```
#Use Describe Package to calculate Descriptive Statistic
df_moneyball_des <- describe(data_train_moneyball, na.rm=TRUE,
interp=FALSE, skew=TRUE, ranges=TRUE, trim=.1, type=3,
check=TRUE, fast=FALSE, quant=c(.1,.25,.75,.90), IQR=TRUE)

# Determine missing value and missing value ratio
df_moneyball_des$missing_values <- df_moneyball_des[1,2] -
df_moneyball_des$n
df_moneyball_des$missing_values_ratio <-
round(df_moneyball_des$missing_values/df_moneyball_des[1,2]*100,
digits = 4)

df_moneyball_des_display <- subset(df_moneyball_des, select =
c(2, 3, 4, 5, 8, 9,  10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20))

knitr::kable(df_moneyball_des_display[-1,])
```

Response variables statistics — TOTAL_WINS

````
knitr::kable(df_moneyball_des_display[2,])
```
h2 <- ggplot(data_train_moneyball, aes(x = TARGET_WINS)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
4)
h2
```


bp2 <- ggplot(data_train_moneyball, aes(x= " ", y = TARGET_WINS))
+
       stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp2

get_outliers <-  function(x, n = 10) {

  v <- abs(x-mean(x,na.rm=TRUE)) > 3*sd(x,na.rm=TRUE)
````

```
  # capture all observations falling into outlier definition sort
descending
  obs <- sort(unique(x[v]), decreasing = T)

  # handle cases where the number of observations is less than
  # the parameter n to return for the top and bottom n values
  if (length(obs) < 2*n) {n <- floor(length(obs)/2)}

  hi <- obs[1:n]

  low <- obs[length(obs):(length(obs)-n +1)]

  # remove dupilcate entries from the lower bound outliers
  low <- setdiff(low, hi)

  return (list(Obs=obs, Hi=hi, Low=low))

}

# this returns a list of vectors; this could be a list of
dataframes if it's easier for output
# Obs = all observations
# Hi = top n observations
# Low = bottom n observations

o2 <- get_outliers(data_train_moneyball$TARGET_WINS)
o2
```

BATTING_H Statistics

```
knitr::kable(df_moneyball_des_display[3,])

```
```{r Batting_h hist, echo=FALSE}

h3 <- ggplot(data_train_moneyball, aes(x = BATTING_H)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
10)
h3
```
```{r Batting_h boxplot, echo=FALSE}

bp3 <- ggplot(data_train_moneyball, aes(x= " ", y = BATTING_H)) +
       stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp3
```
```{r Batting_h Outliers, echo=FALSE}

o3 <- get_outliers(data_train_moneyball$BATTING_H, 10)
o3
```

BATTING_2B Staistics

```
knitr::kable(df_moneyball_des_display[4,])


h4 <- ggplot(data_train_moneyball, aes(x = BATTING_2B)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
5)
h4

bp4 <- ggplot(data_train_moneyball, aes(x= " ", y = BATTING_2B))
+
       stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp4


o4 <- get_outliers(data_train_moneyball$BATTING_2B)
o4
```

BATTING_3B Statics

```
knitr::kable(df_moneyball_des_display[5,])



h5 <- ggplot(data_train_moneyball, aes(x = BATTING_3B)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
5)
h5

bp5 <- ggplot(data_train_moneyball, aes(x= " ", y = BATTING_3B))
+
       stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp5

o5 <- get_outliers(data_train_moneyball$BATTING_3B)
o5
```

BATTING_HR Statics

```
knitr::kable(df_moneyball_des_display[6,])


h6 <- ggplot(data_train_moneyball, aes(x = BATTING_HR)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
5)
h6

bp6 <- ggplot(data_train_moneyball, aes(x= " ", y = BATTING_HR))
+
```

```
        stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp6

o6 <- get_outliers(data_train_moneyball$BATTING_HR)
o6
```

BATTING_BB Statistics

```
knitr::kable(df_moneyball_des_display[7,])


h7 <- ggplot(data_train_moneyball, aes(x = BATTING_BB)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
5)
h7

bp7 <- ggplot(data_train_moneyball, aes(x= " ", y = BATTING_BB))
+
        stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp7
```
```{r Batting_bb Outliers, echo=FALSE}

o7 <- get_outliers(data_train_moneyball$BATTING_BB)
o7
```

BATTING_SO Statistics

```
knitr::kable(df_moneyball_des_display[8,])

h8 <- ggplot(data_train_moneyball, aes(x = BATTING_SO)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
20)
h8

bp8 <- ggplot(data_train_moneyball, aes(x= " ", y = BATTING_SO))
+
        stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp8

o8 <- get_outliers(data_train_moneyball$BATTING_SO)
o8
```

BASERUN_SB Statistics

```
knitr::kable(df_moneyball_des_display[9,])
```

```
h9 <- ggplot(data_train_moneyball, aes(x = BASERUN_SB)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
5)
h9

bp9 <- ggplot(data_train_moneyball, aes(x= " ", y = BASERUN_SB))
+
       stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp9

o9 <- get_outliers(data_train_moneyball$BASERUN_SB)
o9
```

BASERUN_CS Statistics

```
knitr::kable(df_moneyball_des_display[10,])


h10 <- ggplot(data_train_moneyball, aes(x = BASERUN_CS)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
5)
h10


bp10 <- ggplot(data_train_moneyball, aes(x= " ", y = BASERUN_CS))
+
       stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp10

o10 <- get_outliers(data_train_moneyball$BASERUN_CS)
o10
```

BATTING_HBP Statistics

```
knitr::kable(df_moneyball_des_display[11,])


h11 <- ggplot(data_train_moneyball, aes(x = BATTING_HBP)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
3)
h11

bp11 <- ggplot(data_train_moneyball, aes(x= " ", y =
BATTING_HBP)) +
       stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp11
```

```
o11 <- get_outliers(data_train_moneyball$BATTING_HBP)
o11
```

PITCHING_H Statistics

```
knitr::kable(df_moneyball_des_display[12,])



h12 <- ggplot(data_train_moneyball, aes(x = PITCHING_H)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
2)
h12
```

```
```

```
bp12 <- ggplot(data_train_moneyball, aes(x= " ", y = PITCHING_H))
+
       stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp12
```

```
```

```
o12 <- get_outliers(data_train_moneyball$PITCHING_H)
o12
```

PITCHING_HR Statistics

```
knitr::kable(df_moneyball_des_display[13,])
```

```
```

```
h13 <- ggplot(data_train_moneyball, aes(x = PITCHING_HR)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
5)
h13
```

```
```

```
bp13 <- ggplot(data_train_moneyball, aes(x= " ", y =
PITCHING_HR)) +
       stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp13
```

```
```

```
o13 <- get_outliers(data_train_moneyball$PITCHING_HR)
o13
```

PITCHING_BB Statistics

```
knitr::kable(df_moneyball_des_display[14,])
```

```
```

```
h14 <- ggplot(data_train_moneyball, aes(x = PITCHING_BB)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
5)
h14
```

```
```

```
bp14 <- ggplot(data_train_moneyball, aes(x= " ", y =
PITCHING_BB)) +
       stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp14
```

```
o14 <- get_outliers(data_train_moneyball$PITCHING_BB)
o14
```

PITCHING_SO Statistics

```
knitr::kable(df_moneyball_des_display[15,])

```

```
h15 <- ggplot(data_train_moneyball, aes(x = PITCHING_SO)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
5)
h15
```

```
bp15 <- ggplot(data_train_moneyball, aes(x= " ", y =
PITCHING_SO)) +
       stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp15
```

```
o15 <- get_outliers(data_train_moneyball$PITCHING_SO)
o15
```

FIELDING_E Statistics

```
knitr::kable(df_moneyball_des_display[16,])

h16 <- ggplot(data_train_moneyball, aes(x = FIELDING_E)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
3)
h16
```

```
bp16 <- ggplot(data_train_moneyball, aes(x= " ", y = FIELDING_E))
+
       stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp16
```

```
o16 <- get_outliers(data_train_moneyball$FIELDING_E)
```

```
o16
```

FIELDING_DP Statistics

```
knitr::kable(df_moneyball_des_display[17,])

```
```{r fielding_dp hist, echo=FALSE}

h17 <- ggplot(data_train_moneyball, aes(x = FIELDING_DP)) +
geom_histogram(colour = "black", fill = "light blue", binwidth =
3)
h17
```

bp17 <- ggplot(data_train_moneyball, aes(x= " ", y =
FIELDING_DP)) +
       stat_boxplot(geom ='errorbar') +
  geom_boxplot(fill = "light green", outlier.colour = "red",
outlier.shape = 1)
bp17

o17 <- get_outliers(data_train_moneyball$FIELDING_DP)
o17
```

## Missing Values

```
plot_missing(data_train_moneyball)
```

Correlation between variables

```
#correlation plot
par(mfrow=c(1,1))
mb_corr <- dplyr::select(data_train_moneyball, -INDEX)
corrplot(cor(mb_corr, use = "na.or.complete"), order = “hclust")
```

Data Transformation

Replacing remaining 0 values with NA

```
# removal of outliers
data_train_moneyball_transformed <- data_train_moneyball %>%
  filter(TARGET_WINS >=22 & TARGET_WINS <= 124 & BATTING_H <=
1876 & BATTING_2B >= 116 & BATTING_2B <= 376 & BATTING_BB >= 292
&
           BATTING_BB <= 879 & BATTING_SO >= 326 & BATTING_SO <=
1535 & PITCHING_HR <= 258 & PITCHING_SO <= 1450 & BATTING_3B>=11
& BATTING_3B<=153 & PITCHING_H <= 3000)

# removal of discarded predictors
data_train_moneyball_transformed <- dplyr::select
(data_train_moneyball_transformed, -BATTING_HBP, -BASERUN_CS, -
INDEX)
```
```

```
nrow(data_train_moneyball)-nrow(data_train_moneyball_transformed)

data_train_moneyball_transformed[data_train_moneyball_transformed
== 0] <- NA
```

Adding BsR

```
#Adding BATTING_1B
data_train_moneyball_transformed$BATTING_1B <-
data_train_moneyball_transformed$BATTING_H -
(data_train_moneyball_transformed$BATTING_2B +
data_train_moneyball_transformed$BATTING_3B +
data_train_moneyball_transformed$BATTING_HR)

#Adding WHGP, PTICHING_SO_BB, and BATTING_BB_SO
data_train_moneyball_transformed <-
data_train_moneyball_transformed %>%
  mutate(BATTING_TB = (BATTING_1B + BATTING_2B * 2 + BATTING_3B *
3 + BATTING_HR * 4)) %>%
  mutate(WHGP = (PITCHING_H + PITCHING_BB)/162) %>%
  mutate(PITCHING_SO_BB = PITCHING_SO/PITCHING_BB) %>%
  mutate(BATTING_BB_SO = BATTING_BB/BATTING_SO)

# Determining Avg AB
#Baseball Reference data
AVG <-
read.csv("https://raw.githubusercontent.com/bkreis84/Business-
Analytics/master/HW1/Baseball%20Reference%20AVG.csv")
avgAB <- mean(AVG$AB, na.rm = TRUE)*162

# Deriving BSR
data_train_moneyball_transformed <-
data_train_moneyball_transformed %>%
  mutate(BsR = (((BATTING_H + BATTING_BB - BATTING_HR) *
((1.4*BATTING_TB - .6*BATTING_H -3*BATTING_HR
+0.1*BATTING_BB)*1.02)) /
                  (((1.4*BATTING_TB - .6*BATTING_H -3*BATTING_HR
+.1*BATTING_BB)*1.02) + avgAB - BATTING_H)) + BATTING_HR)
#Dropping BATTING_H
data_train_moneyball_transformed <-
dplyr::select(data_train_moneyball_transformed, -BATTING_H)
```

Addressing Skewness of Some Variables with Box-Cox

```
#set.seed(4234)
#g1 <- BoxCoxTrans(data_train_moneyball_transformed$TARGET_WINS, na.rm
= TRUE)
#g2 <- BoxCoxTrans(data_train_moneyball_transformed$BATTING_2B, na.rm
= TRUE)
#g3 <- BoxCoxTrans(data_train_moneyball_transformed$BATTING_3B, na.rm
= TRUE)
#g4 <- BoxCoxTrans(data_train_moneyball_transformed$BATTING_HR, na.rm
= TRUE)
```

```
#g5 <- BoxCoxTrans(data_train_moneyball_transformed$BATTING_BB, na.rm
= TRUE)
#g6 <- BoxCoxTrans(data_train_moneyball_transformed$BATTING_SO, na.rm
= TRUE)
#g7 <- BoxCoxTrans(data_train_moneyball_transformed$BASERUN_SB, na.rm
= TRUE)
#g8 <- BoxCoxTrans(data_train_moneyball_transformed$PITCHING_H, na.rm
= TRUE)
#g9 <- BoxCoxTrans(data_train_moneyball_transformed$PITCHING_HR, na.rm
= TRUE)
#g10 <- BoxCoxTrans(data_train_moneyball_transformed$PITCHING_BB,
na.rm = TRUE)
#g11 <- BoxCoxTrans(data_train_moneyball_transformed$PITCHING_SO,
na.rm = TRUE)
#g12 <- BoxCoxTrans(data_train_moneyball_transformed$FIELDING_E, na.rm
= TRUE)
#g13 <- BoxCoxTrans(data_train_moneyball_transformed$FIELDING_DP,
na.rm = TRUE)
#g14 <- BoxCoxTrans(data_train_moneyball_transformed$BATTING_1B, na.rm
= TRUE)
#g15 <- BoxCoxTrans(data_train_moneyball_transformed$BATTING_TB, na.rm
= TRUE)
#g16 <- BoxCoxTrans(data_train_moneyball_transformed$WHGP, na.rm =
TRUE)
#g17 <- BoxCoxTrans(data_train_moneyball_transformed$PITCHING_SO_BB,
na.rm = TRUE)
#g18 <- BoxCoxTrans(data_train_moneyball_transformed$BATTING_BB_SO,
na.rm = TRUE)
#g19 <- BoxCoxTrans(data_train_moneyball_transformed$BsR, na.rm =
TRUE)

#lambdas <- c(g2$lambda, g3$lambda, g4$lambda, g5$lambda, g6$lambda,
g7$lambda, g8$lambda, g9$lambda, g10$lambda, g11$lambda, g12$lambda,
g13$lambda, g14$lambda, g15$lambda, g16$lambda, g17$lambda,
g18$lambda, g19$lambda)


#trans_bc <- preProcess(data_train_moneyball_transformed, method =
"BoxCox")

#data_train_moneyball_transformed_2 <- predict(trans_bc,
data_train_moneyball_transformed)

# if no box cox transformation to stream line further code
data_train_moneyball_transformed_2 <- data_train_moneyball_transformed
```

Imputation of Missing Values

```
#If Box Cox has not been applied
md.pattern(data_train_moneyball_transformed_2)

#Visualize missing values with VIM
```

```
aggr_plot <- aggr(data_train_moneyball_transformed_2,
                  col=c('navyblue','red'), numbers=TRUE,
sortVars=TRUE, labels=names(data_train_moneyball_transformed_2),
cex.axis=.7, gap=3, ylab=c("Histogram of missing data”,"Pattern"))

#Imput missing data using mice
data_train_moneyball_imput_pmm <-
mice(data_train_moneyball_transformed_2,m=5,maxit=50,meth='pmm',seed=5
00)

data_train_moneyball_imput_nboot <-
mice(data_train_moneyball_transformed_2,m=5,maxit=50,meth='norm.boot',
seed=500)

data_train_moneyball_imput_rf <-
mice(data_train_moneyball_transformed_2,m=5,maxit=50, meth='rf',
seed=500)

data_train_moneyball_imput <-
mice(data_train_moneyball_transformed_2,m=5,maxit=50, seed=500)
```
```
#summary(data_train_moneyball_imput)
#summary(data_train_moneyball_imput_pmm)
#summary(data_train_moneyball_imput_nboot)
```
```
#data_train_moneyball_imput$imp$FIELDING_DP

#data_train_moneyball_imput$imp$BASERUN_SB

#Density plot of values imputed
densityplot(data_train_moneyball_imput)

densityplot(data_train_moneyball_imput_pmm)

densityplot(data_train_moneyball_imput_nboot)

densityplot(data_train_moneyball_imput_rf)

stripplot(data_train_moneyball_imput_nboot, pch = 20, cex = 1.2)
```
```
# replacing missing values with imputed values from 2nd data set
data_train_moneyball_complete <-
complete(data_train_moneyball_imput_nboot,2)
```

Transformation Recap

```
knitr::kable(describe(data_train_moneyball_complete))

ggplot(stack(data_train_moneyball_complete), aes(values))+
  facet_wrap(~ind, scales = "free") +
  geom_histogram(fill = "light blue", colour="black") +
  theme(legend.position="none")
```

~~~
```
ggplot(stack(data_train_moneyball_complete), aes(x = ind, y = values,
fill=ind))+
  facet_wrap(~ind, scales = "free") +
  geom_boxplot() +
  theme(legend.position="none")

write.csv(data_train_moneyball_complete, file = "data_group2_nbc.csv")
~~~

Build Models

Model 1 - Base model with backward selection

```
backReg <-
read.csv("https://raw.githubusercontent.com/vbriot28/Data621_group2/ma
ster/data_group2_nbc.csv")

boxBackReg <-
read.csv("https://raw.githubusercontent.com/vbriot28/Data621_group2/ma
ster/data_group2.csv")

library(stats)

#TARGET_WINS as the dependent variable all predictors non-transformed
data

mod1 <- step(lm(TARGET_WINS ~ PITCHING_H + PITCHING_HR + PITCHING_BB +
PITCHING_SO +
            PITCHING_SO_BB + BATTING_2B + BATTING_3B + BATTING_HR +
BATTING_BB +
         BATTING_SO + BATTING_1B + BATTING_TB + BATTING_BB_SO +
            FIELDING_E + FIELDING_DP + BASERUN_SB + BsR,
        data = backReg),
     direction = "backward")
summary(mod1)

#Removal of BSR, BATTING_1B and PRITCHING_SO based on AIC values -
non-transformed data
mod2 <- step(lm(TARGET_WINS ~ PITCHING_H + PITCHING_HR + PITCHING_BB +
            PITCHING_SO_BB + BATTING_2B + BATTING_3B + BATTING_HR +
BATTING_BB +
         BATTING_SO + BATTING_TB + BATTING_BB_SO +
            FIELDING_E + FIELDING_DP + BASERUN_SB,
        data = backReg),
     direction = "backward")

summary(mod2)

#Removal of BATTING_2B - non-transformed data
mod3 <- step(lm(TARGET_WINS ~ PITCHING_H + PITCHING_HR + PITCHING_BB +
            PITCHING_SO_BB + BATTING_3B + BATTING_HR + BATTING_BB +
         BATTING_SO + BATTING_TB + BATTING_BB_SO +
            FIELDING_E + FIELDING_DP + BASERUN_SB,
```

```
        data = backReg),
     direction = "backward")

summary(mod3)

#Removal of PITCHING_SO_BB and BATTING_BB_SO - non-transformed data
mod4 <- step(lm(TARGET_WINS ~ PITCHING_H + PITCHING_HR + PITCHING_BB +
            BATTING_3B + BATTING_HR + BATTING_BB + BATTING_SO +
BATTING_TB +
            FIELDING_E + FIELDING_DP + BASERUN_SB,
        data = backReg),
     direction = "backward")

summary(mod4)
```

Evaluate the Model

```
library(car)

residualPlots(mod3)
qqPlot(mod3, id.n=3)
outlierTest(mod3)
influenceIndexPlot(mod3, id.n=3)
influencePlot(mod3, id.n=3)
hist(mod3$res)
```

Model 2 - Total Base Model with forward selection

```
model2 <- step(lm(TARGET_WINS ~ BATTING_TB, data =
data_train_moneyball_complete),
               direction = "forward",
               scope  = ~ BATTING_1B +
                 BATTING_2B +
                 BATTING_3B +
                 BATTING_HR +
                 BATTING_BB +
                 BATTING_SO +
                 BASERUN_SB +
                 PITCHING_H +
                 PITCHING_HR +
                 PITCHING_BB +
                 PITCHING_SO +
                 FIELDING_E +
                 FIELDING_DP +
                 BATTING_TB)

summary(model2)

residualPlots(model2)
qqPlot(model2, id.n=3)
outlierTest(model2)
influenceIndexPlot(model2, id.n=3)
influencePlot(model2, id.n=3)
```

```
hist(model2$res, main="Histogram of Residuals”)
```

Model 3 - Walks and Hits Per Game Played (WHGP)

```
library(MASS)
library(car)

#mbstats <-
read.csv(“https://raw.githubusercontent.com/vbriot28/Data621_group2/ma
ster/data_group2_nbc.csv")

model3 <- step(lm(TARGET_WINS ~ WHGP, data =
data_train_moneyball_complete), direction="forward",
                 scope= ~ BATTING_TB  +   BATTING_BB  +  BATTING_SO  +
                          BASERUN_SB   +  PITCHING_H  +  PITCHING_HR +
PITCHING_BB  +
                          PITCHING_SO  +  FIELDING_E  +  FIELDING_DP
+ WHGP)

formula(model3)
```

Model Summary

```
model3 <- update(model3, . ~ . - BATTING_TB)

summary(model3)

extractAIC(model3)


formula(model3)
```

Model Diagnostics

```
residualPlots(model3)

qqPlot(model3, id.n=3, main="Q-Q Plot”)
```

Outliers & Influence Points

```
influenceIndexPlot(model3, id.n=3)


influencePlot(model3, id.n=3)


hist(model3$residuals, main="Histogram of Residuals”)
```

Model 4

```
#data <-
read.csv("https://raw.githubusercontent.com/vbriot28/Data621_group2/ma
ster/data_group2_nbc.csv")

model4 <- step(lm(TARGET_WINS ~ BsR, data =
data_train_moneyball_complete),
                  direction = "forward",
```

```
               scope = ~ BsR + BATTING_2B + BATTING_3B + BATTING_HR +
BATTING_BB + BATTING_SO +
                 BASERUN_SB + PITCHING_H + PITCHING_HR + PITCHING_BB +
PITCHING_SO +
                 FIELDING_E + FIELDING_DP + BATTING_1B + BATTING_TB +
WHGP + PITCHING_SO_BB + BATTING_BB_SO
)


#tbl4 <- tidy(model4)
#kable(tbl4)
#kable(glance(model4))
summary(model4)

#Run Bi-directional model for Model 4
model4b <- step(lm(TARGET_WINS ~ BsR, data =
data_train_moneyball_complete),
               direction = "both",
               scope = ~ BsR + BATTING_2B + BATTING_3B + BATTING_HR +
BATTING_BB + BATTING_SO +
                 BASERUN_SB + PITCHING_H + PITCHING_HR + PITCHING_BB +
PITCHING_SO +
                 FIELDING_E + FIELDING_DP + BATTING_1B + BATTING_TB +
WHGP + PITCHING_SO_BB + BATTING_BB_SO
)


#tbl4b <- tidy(model4b)
#kable(tbl4b)
kable(glance(model4b))

summary(model4b)

residualPlots(model4b)
qqPlot(model4b, id.n=3, main="Q-Q Plot")
influenceIndexPlot(model4b, id.n=3)
influencePlot(model4b, id.n=3)
hist(model4b$residuals, main="Histogram of Residuals")
```

Evaluating models & Selecting best one

```
### Calculate Accuracy, based on this tutorial:
http://rstatistics.net/robust-regression/

fit.Predicted <- predict(model1_3, model1_3$model)
fit.Actuals.pred <- cbind(fit.Predicted, model1_3$model[1])
accuracy1 <- round(mean(apply(fit.Actuals.pred, 1, min)/
apply(fit.Actuals.pred, 1, max)),3)

fit.Predicted <- predict(model2, model2$model)
fit.Actuals.pred <- cbind(fit.Predicted, model2$model[1])
```

```
accuracy2 <- round(mean(apply(fit.Actuals.pred, 1, min)/
apply(fit.Actuals.pred, 1, max)),3)

fit.Predicted <- predict(model3, model3$model)
fit.Actuals.pred <- cbind(fit.Predicted, model3$model[1])
accuracy3 <- round(mean(apply(fit.Actuals.pred, 1, min)/
apply(fit.Actuals.pred, 1, max)),3)

fit.Predicted <- predict(model4b, model4b$model)
fit.Actuals.pred <- cbind(fit.Predicted, model4b$model[1])
accuracy4b <- round(mean(apply(fit.Actuals.pred, 1, min)/
apply(fit.Actuals.pred, 1, max)),3)

#anova(model1_3, model4b)
```

Using our model to make prediction

```
data_evaluation_moneyball <-
read.csv("https://raw.githubusercontent.com/vbriot28/Data621_grou
p2/master/moneyball-evaluation-data.csv", header = TRUE)

colnames(data_evaluation_moneyball) = gsub("TEAM_", "",
colnames(data_evaluation_moneyball))

# removal of outliers
data_evaluation_moneyball_transformed <-
data_evaluation_moneyball %>%
  filter(BATTING_H <= 1876 & BATTING_2B >= 116 & BATTING_2B <=
376 & BATTING_BB >= 292 &
           BATTING_BB <= 879 & BATTING_SO >= 326 & BATTING_SO <=
1535 & PITCHING_HR <= 258 & PITCHING_SO <= 1450 & BATTING_3B>=11
& BATTING_3B<=153 & PITCHING_H <= 3000)


# removal of discarded predictors
data_evaluation_moneyball_transformed <- dplyr::select
(data_evaluation_moneyball_transformed, -BATTING_HBP, -
BASERUN_CS)


dropped_rows_evaluation <- nrow(data_evaluation_moneyball)-
nrow(data_evaluation_moneyball_transformed)

data_evaluation_moneyball_transformed[data_evaluation_moneyball_t
ransformed == 0] <- NA


#Adding BATTING_1B
data_evaluation_moneyball_transformed$BATTING_1B <-
data_evaluation_moneyball_transformed$BATTING_H -
(data_evaluation_moneyball_transformed$BATTING_2B +
```

```
data_evaluation_moneyball_transformed$BATTING_3B +
data_evaluation_moneyball_transformed$BATTING_HR)

#Adding WHGP, PTICHING_SO_BB, and BATTING_BB_SO
data_evaluation_moneyball_transformed <-
data_evaluation_moneyball_transformed %>%
  mutate(BATTING_TB = (BATTING_1B + BATTING_2B * 2 + BATTING_3B *
3 + BATTING_HR * 4)) %>%
  mutate(WHGP = (PITCHING_H + PITCHING_BB)/162) %>%
  mutate(PITCHING_SO_BB = PITCHING_SO/PITCHING_BB) %>%
  mutate(BATTING_BB_SO = BATTING_BB/BATTING_SO)

# Deriving BSR
data_evaluation_moneyball_transformed <-
data_evaluation_moneyball_transformed %>%
  mutate(BsR = (((BATTING_H + BATTING_BB - BATTING_HR) *
((1.4*BATTING_TB - .6*BATTING_H -3*BATTING_HR
+0.1*BATTING_BB)*1.02)) /
                  (((1.4*BATTING_TB - .6*BATTING_H -3*BATTING_HR
+.1*BATTING_BB)*1.02) + avgAB - BATTING_H)) + BATTING_HR)
#Dropping BATTING_H
data_evaluation_moneyball_transformed <-
dplyr::select(data_evaluation_moneyball_transformed, -BATTING_H)

#impute missing value
data_evaluation_moneyball_rf <-
mice(data_evaluation_moneyball_transformed,m=5,maxit=50,
meth='rf', seed=500)
data_evaluation_moneyball_complete <-
complete(data_evaluation_moneyball_rf,2)

write.csv(data_evaluation_moneyball_complete, file =
"data_group2_evaluation_nbc.csv")

` ` `
` ` `{r predict TARGET_WINS, echo=FALSE}
#Load transformed data set for replicability of results
data_evaluation_moneyball_predict <-
read.csv("https://raw.githubusercontent.com/vbriot28/Data621_grou
p2/master/data_group2_evaluation_nbc.csv", header = TRUE)

fit.Predicted1 <- predict(model1_3,
data_evaluation_moneyball_predict)

fit.Predicted1_rd <- round(fit.Predicted1,0)

#write all prediction to file
write.csv(as.data.frame(fit.Predicted1_rd), file =
"data_group2_prediction_model1.csv")

#Basic stats across all data sets
summary(data_train_moneyball$TARGET_WINS)
```

```
summary(data_train_moneyball_complete$TARGET_WINS)
summary(fit.Predicted1_rd)
```