Chelsea Clinger

411 Generalized Linear Methods 56

Unit 1

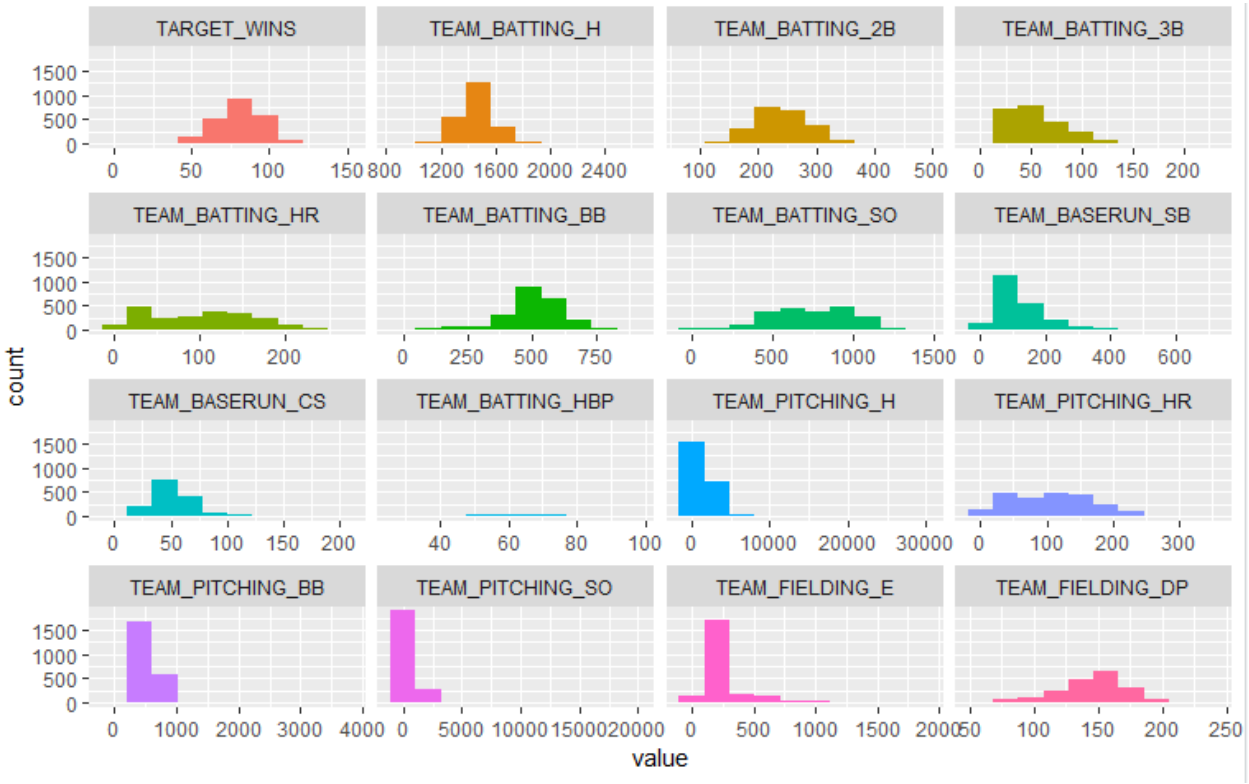
Introduction

The purpose of this assignment is to analyze data produced from baseball games to attempt to predict the number of wins a team will have in the upcoming season.

The analysis conducted in this assignment utilizes various statistics observed in baseball games. The dataset analyzed includes 17 statistics created by actions at bat by the players. These result in 2,276 observations. In any spot, some actions taken by a player are beneficial and can lead to additional points or even a win, while some actions can cause the other team to score or even lose a game.

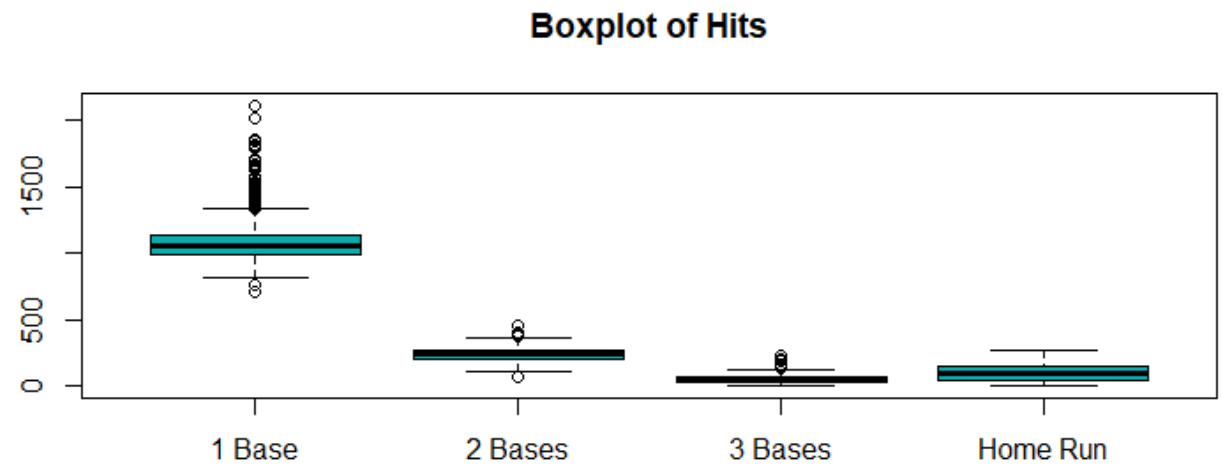
Section 1 – Data Exploration

Before beginning any analysis on the variables, an initial view into each of them will be conducted. The first histogram below illustrates the distribution of target wins, we can see that most teams have between 70 and 90 target wins this figure is supported by some logical thinking. If there are 160 games in a typical season, we would expect teams to win/lose half of the time, resulting in around 80 wins per team.

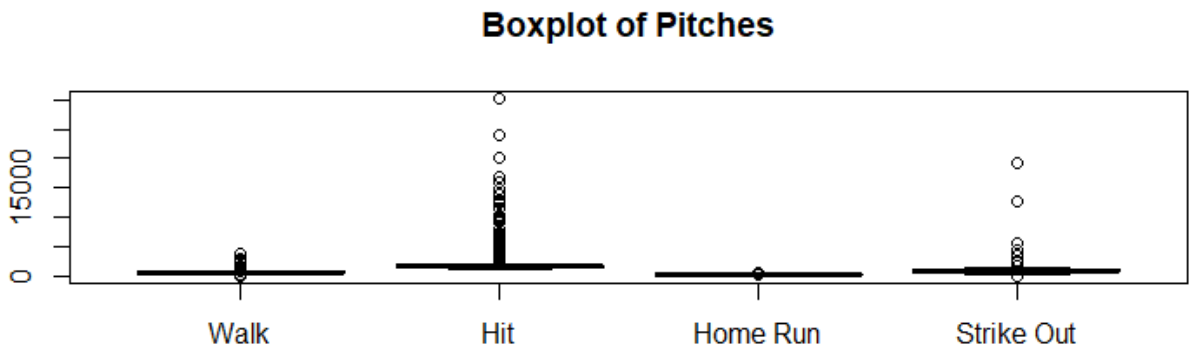


The variables can be split up into multiple categories; Batting, Fielding and Pitching. Seven of the variables are for the outcome of the time spent at bat by a player. A quick glance at each of the seven variable distributions allows us to observe that while there are a lot of hits at bat (typically between

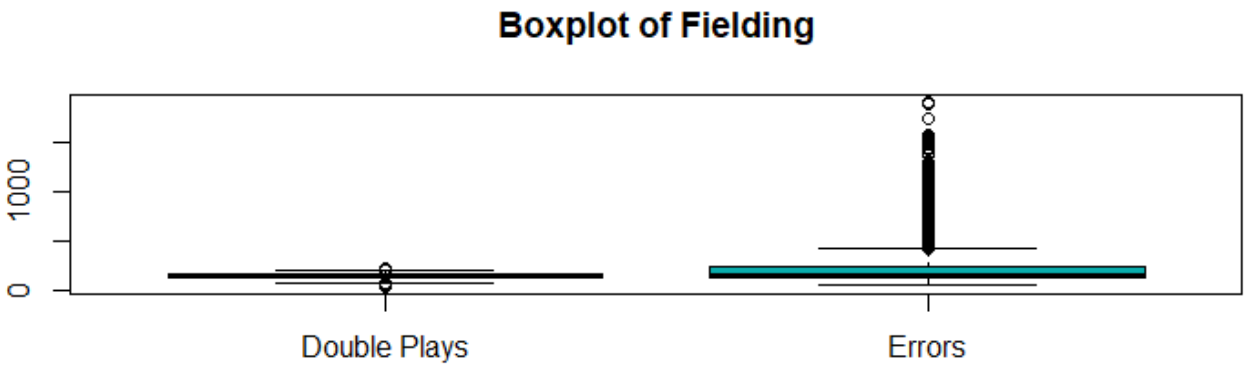
1,383 and 1,537) the number decreases significantly from double to triple to homerun. When a batter makes a hit, the majority result in one base which is evident below. Running two or three bases is significantly less likely. However, what is somewhat surprising is that home runs occur more often than running and stopping at third.



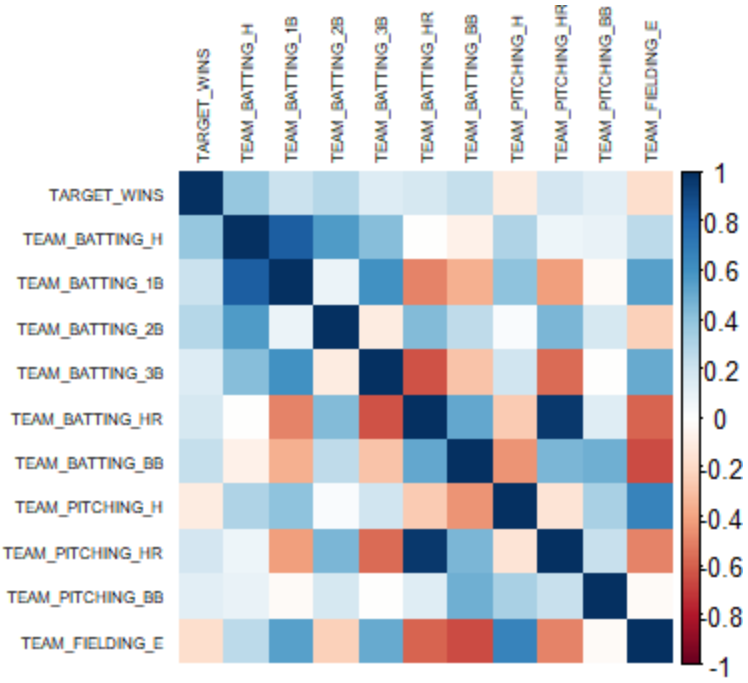
Since we have considered what the outcome is when a player makes a hit, next we will observe the data available for pitching the ball. The below graph is extremely hard to observe the ranges for. This makes a very distinguishable observation that there are far more hits than any other scenario.



Lastly there are the field positions. The two outcomes presented in the data include double plays and errors. There is a much more significant volume of errors experienced than double plays in the graph below.



The graph on the right illustrates how the variables are correlated with each other. Initially it is beneficial to observe the relationship between target wins and the other variables to see what drives the wins variable in a positive or negative way. The variable that appears to have the largest positive influence on target wins is the number of base hits by batters. Thus the initial most apparent way to win a game is to get any number of bases provided the batter hits the ball. While that fact is not at all surprising, what goes against initial thinking is that there is a larger correlation between walking and wins than there is between home runs and wins, indicating that it is more influential to the overall number of target wins to walk at bat than it is to score a home run.



An initial overview of the data indicates that there are many missing values in the data set. Six variables have missing data in them and will need to be considered in future analysis. This will be especially concerning for the Batters hit by pitch since 91% of the values are blank.

	variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
1	INDEX	0	0.00	0	0.00	0	0	integer	2276
2	TARGET_WINS	1	0.04	0	0.00	0	0	integer	108
3	TEAM_BATTING_H	0	0.00	0	0.00	0	0	integer	569
4	TEAM_BATTING_2B	0	0.00	0	0.00	0	0	integer	240
5	TEAM_BATTING_3B	2	0.09	0	0.00	0	0	integer	144
6	TEAM_BATTING_HR	15	0.66	0	0.00	0	0	integer	243
7	TEAM_BATTING_BB	1	0.04	0	0.00	0	0	integer	533
8	TEAM_BATTING_SO	20	0.88	102	4.48	0	0	integer	822
9	TEAM_BASERUN_SB	2	0.09	131	5.76	0	0	integer	348
10	TEAM_BASERUN_CS	1	0.04	772	33.92	0	0	integer	128
11	TEAM_BATTING_HBP	0	0.00	2085	91.61	0	0	integer	55
12	TEAM_PITCHING_H	0	0.00	0	0.00	0	0	integer	843
13	TEAM_PITCHING_HR	15	0.66	0	0.00	0	0	integer	256
14	TEAM_PITCHING_BB	1	0.04	0	0.00	0	0	integer	535
15	TEAM_PITCHING_SO	20	0.88	102	4.48	0	0	integer	823
16	TEAM_FIELDING_E	0	0.00	0	0.00	0	0	integer	549
17	TEAM_FIELDING_DP	0	0.00	286	12.57	0	0	integer	144

Section 2 – Data Preparation

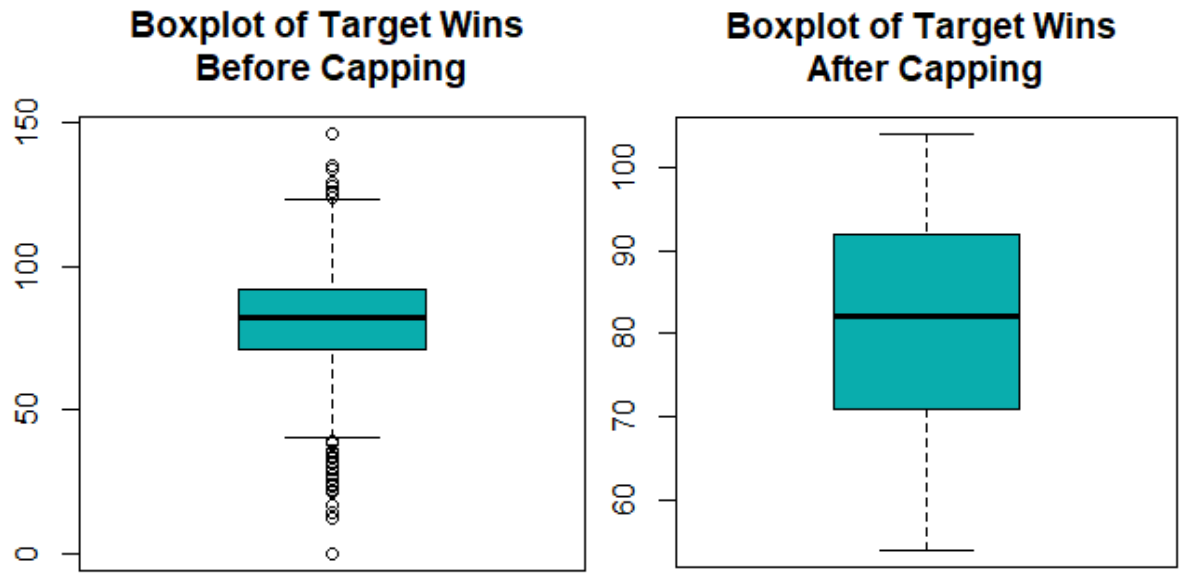
Fixing missing values in data sets can be complicated, it is not feasible in many cases, such as this one, to remove entire data entries because of missing data. Therefore there are many other methods of resolving issues with missing data other than simply removing data like utilizing the data set mean, median or mode. For the purposes of this assignment I utilized the knnImputation function. This

function utilizes the nearest neighbor methodology to replace missing values. The new values are calculated utilizing the weighted average of the nearest neighbors.

	variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
1	INDEX	0	0.00	0	0	0	0	integer	2276
2	TARGET_WINS	1	0.04	0	0	0	0	integer	108
3	TEAM_BATTING_H	0	0.00	0	0	0	0	integer	569
4	TEAM_BATTING_2B	0	0.00	0	0	0	0	integer	240
5	TEAM_BATTING_3B	2	0.09	0	0	0	0	integer	144
6	TEAM_BATTING_HR	15	0.66	0	0	0	0	integer	243
7	TEAM_BATTING_BB	1	0.04	0	0	0	0	integer	533
8	TEAM_BATTING_SO	20	0.88	0	0	0	0	numeric	924
9	TEAM_BASERUN_SB	2	0.09	0	0	0	0	numeric	479
10	TEAM_BASERUN_CS	1	0.04	0	0	0	0	numeric	900
11	TEAM_BATTING_HBP	0	0.00	0	0	0	0	numeric	2140
12	TEAM_PITCHING_H	0	0.00	0	0	0	0	integer	843
13	TEAM_PITCHING_HR	15	0.66	0	0	0	0	integer	256
14	TEAM_PITCHING_BB	1	0.04	0	0	0	0	integer	535
15	TEAM_PITCHING_SO	20	0.88	0	0	0	0	numeric	925
16	TEAM_FIELDING_E	0	0.00	0	0	0	0	integer	549
17	TEAM_FIELDING_DP	0	0.00	0	0	0	0	numeric	430
18	TEAM_BATTING_1B	0	0.00	0	0	0	0	integer	497

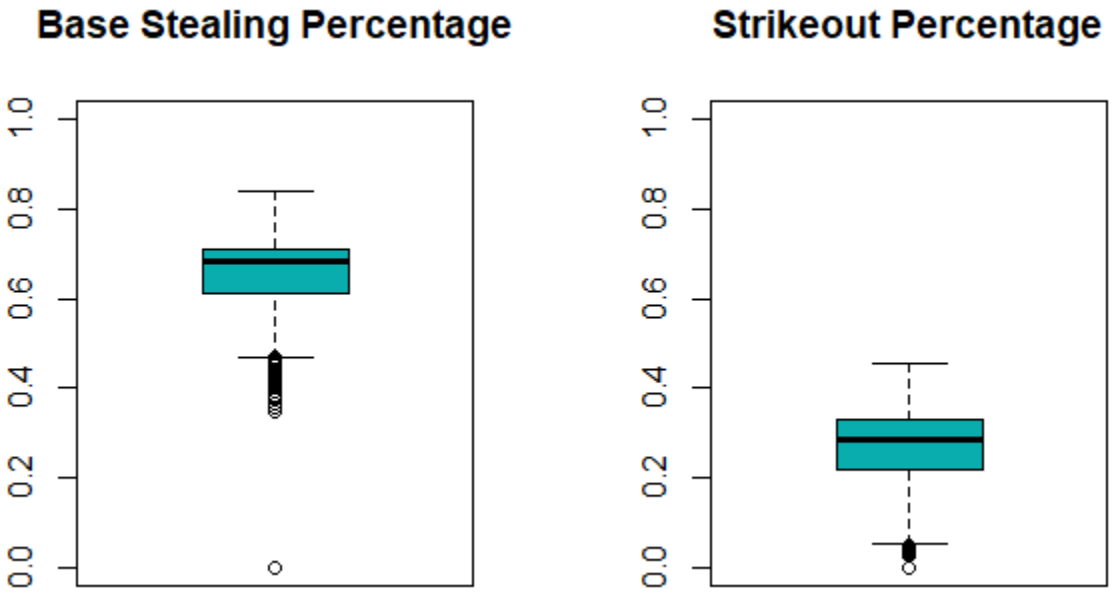
A quick check of the new dataset indicates that there are no longer any missing values and that none of the missing values have been arbitrarily replaced with “0”.

The dataset also has a number of outliers that need to be remediated. When looking at the response variable, Target Wins, it is clear that there are a large amount of outliers extending beyond the whiskers of the below left boxplot. These outliers will sway the results of future predictions made by the model. To accommodate for this influence, the target wins variable was “capped” at 5% and 95%. At 5% the value is 54 and at 95% the value is 104. Therefore, values that are below 54 will be replaced with 54 and values that are above 104 will be replaced with 104.



In addition to capping the target wins variable, additional variables have also been made.

- Base Stealing Percentage: How often when it is attempted to steal a base is the base successfully stolen?
- Strikeout Percentage: How often to pitchers strike out the batter?



Roughly 65% of the time a batter attempts to steal a base they are successful however there are some teams that are never successful at stealing bases.

Roughly 27% of pitches result in a strikeout, however at least one pitcher averages close to 50% strikeouts.

Section 3 – Model Building

The first model created to analyze the moneyball data will be utilizing manually selected predictor variables. The variables selected include hits at bat, balls at bat, base stealing percentage, strikeout pitching percentage, home runs pitched and fielding errors.

Utilizing the model output below it is apparent that the pitching variables have a lesser significance in the model while the fielding and batting variables have the most. The adjusted R-Squared for this model is only .2344 which is not a large R-Squared but is a suitable starting point.

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.4909553  4.7460941   1.368    0.172
TEAM_BATTING_H   0.0381966  0.0023717  16.105  < 2e-16 ***
TEAM_BATTING_BB  0.0132636  0.0028366   4.676 3.10e-06 ***
SB_PCT          25.1633252  2.5381540   9.914  < 2e-16 ***
SO_PCT          -6.8568566  5.1670047  -1.327    0.185
TEAM_PITCHING_H  0.0003758  0.0002714   1.385    0.166
TEAM_FIELDING_E -0.0164163  0.0021996  -7.463 1.19e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.1 on 2268 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.2361,    Adjusted R-squared:  0.2341
F-statistic: 116.8 on 6 and 2268 DF,  p-value: < 2.2e-16
```

Some of the most significant variables have a positive relationship with target wins while some have a negative relationship. The base-stealing percentage has a very strong positive relationship with target wins indicating that stealing a base has a greater correlation with winning than even getting to a base (although you would have to in order to steal so that assumption is already made). Additionally fielding errors have a rather significant negative correlation indicating that a significant ways to prevent a loss is to prevent errors.

The second model created utilizing the moneyball data will be created utilizing the stepwise method. This method of regression adds and removes predictor variables from the model. This method starts with no predictors but then adds the most valuable predictors. Once new variables are added the variables that no longer improve the model are removed. The predictor variables selected for this initial model include all hits, singles, doubles, triples, homeruns, and walks for batters, base stealing percentage, strikeout percentage, fielding errors and double plays.

The initial AIC for this model before performing stepwise is 11230.66. The first predictor variable to be removed is the log taken of batting hits which reduces the AIC to 11228.72. Next batting strikeouts, pitching home runs, getting caught stealing, and pitching a walk are all removed to have a final AIC of 11222.57.

```
                    Df Sum of Sq    RSS   AIC
<none>                           311907 11223
+ TEAM_PITCHING_BB    1    142.5 311765 11224
- TEAM_BATTING_2B     1    412.7 312320 11224
+ TEAM_PITCHING_HR    1     98.3 311809 11224
+ TEAM_BASERUN_CS     1     83.5 311824 11224
+ log_TEAM_BATTING_H  1     20.8 311886 11224
+ TEAM_BATTING_SO     1     18.0 311889 11224
- TEAM_BASERUN_SB     1   1988.2 313895 11235
- TEAM_BATTING_BB     1   2027.8 313935 11235
- TEAM_PITCHING_SO    1   2226.7 314134 11237
- SB_PCT              1   3046.8 314954 11243
- SO_PCT              1   3701.6 315609 11247
- TEAM_FIELDING_E     1   7490.2 319397 11275
- TEAM_BATTING_3B     1   8098.6 320006 11279
- TEAM_BATTING_1B     1  10411.2 322318 11295
- TEAM_FIELDING_DP    1  11527.5 323435 11303
- TEAM_BATTING_HR     1  26328.3 338235 11405
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.714e+01  6.154e+00   6.035 1.86e-09 ***
TEAM_BATTING_1B  3.014e-02  3.467e-03   8.693  < 2e-16 ***
TEAM_BATTING_2B  1.153e-02  6.659e-03   1.731 0.083610 .
TEAM_BATTING_3B  1.056e-01  1.378e-02   7.667 2.59e-14 ***
TEAM_BATTING_HR  1.000e-01  7.237e-03  13.824  < 2e-16 ***
TEAM_BATTING_BB  1.211e-02  3.156e-03   3.837 0.000128 ***
TEAM_BASERUN_SB  1.575e-02  4.145e-03   3.799 0.000149 ***
TEAM_PITCHING_SO 2.204e-03  5.482e-04   4.020 6.00e-05 ***
TEAM_FIELDING_E -1.405e-02  1.906e-03  -7.373 2.32e-13 ***
TEAM_FIELDING_DP -1.054e-01  1.153e-02  -9.147  < 2e-16 ***
SB_PCT           1.759e+01  3.739e+00   4.703 2.72e-06 ***
SO_PCT          -3.533e+01  6.816e+00  -5.183 2.37e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.74 on 2264 degrees of freedom
Multiple R-squared:  0.2836,    Adjusted R-squared:  0.2801
F-statistic: 81.49 on 11 and 2264 DF,  p-value: < 2.2e-16
```

The final model created using the stepwise method has all significant variables except for 2nd base and getting caught stealing. The model created after this transformation has an adjusted R-Squared value of .2796 which is slightly higher than the model created manually indicating that it may be a better model.

The regbest function looks at the model's R-Squared and pvalue to determine which variables to include. Utilizing the below graph it can be determined that 9 variables will be selected due to the high R-Squared and the low pvalue.

```
                                    R2        Pvalue
Model with 1 variable    0.07259187  3.781480e-39
Model with 2 variables   0.14124937  6.917073e-76
Model with 3 variables   0.18112279  4.233248e-98
Model with 4 variables   0.21866418 5.175569e-120
Model with 5 variables   0.24778318 1.591382e-137
Model with 6 variables   0.25651756 3.861913e-142
Model with 7 variables   0.26851907 4.987566e-149
Model with 8 variables   0.27497845 2.606751e-152
Model with 9 variables   0.27676445 1.746001e-152
Model with 10 variables 0.27800070 2.587276e-152
Model with 11 variables 0.27864853 9.072770e-152
```

The final model to be created will utilize the reg best function in R. The model created with this function includes each base run including home runs, pitching balls, fielding errors or double plays as well as percentages for strikeouts and base stealing.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     40.773506   5.334816   7.643 3.11e-14 ***
TEAM_BATTING_1B  0.028657   0.003436   8.341  < 2e-16 ***
TEAM_BATTING_2B  0.015452   0.006532   2.366   0.0181 *
TEAM_BATTING_3B  0.104488   0.013440   7.775 1.14e-14 ***
TEAM_BATTING_HR  0.100117   0.006921  14.465  < 2e-16 ***
TEAM_PITCHING_BB 0.006584   0.001563   4.213 2.62e-05 ***
TEAM_FIELDING_E -0.017141   0.001505 -11.387  < 2e-16 ***
TEAM_FIELDING_DP -0.100115  0.011523  -8.689  < 2e-16 ***
SB_PCT          20.865228   2.584420   8.073 1.10e-15 ***
SO_PCT         -32.327478   6.116124  -5.286 1.37e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.79 on 2266 degrees of freedom
Multiple R-squared:  0.2768,    Adjusted R-squared:  0.2739
F-statistic: 96.35 on 9 and 2266 DF,  p-value: < 2.2e-16
```

Each of the variables is considered to be significant except for running 2 bases. This could have some influence by the third baseman assisting to prevent from stealing 3rd. One point that stands out to be is the negative estimate for pitching strikeouts, that seems counter intuitive as strikeouts should have a positive impact on wins. It is also worth noting the very high figure for the base stealing percentage. The adjusted R-Squared for this model is only .2739 which is not a monumental figure but comparable to the other models.

Section 4 – Model Selection

Three models have been created for the money ball data set; a manually selected model, a model with variables selected via stepwise methods and a model created utilizing the regmod function.

The adjusted R-Squared of each model will be utilized to compare the percent of variability of the mean each model represents. Additionally the MSE of each model will be compared to showcase the estimated errors in each model.

	MANUAL	STEPWISE	REGMOD

R-SQUARED	.2341	.2801	.2739
MSE	147.5745	137.1515	138.8041

Due to the stepwise model having both the smallest adjusted r-squared and the smallest errors, this model will be selected going forward.

Conclusion

Three regression models were created to attempt to predict the number of wins a team will achieve in a season. These models included a manual model, stepwise variable selection and the regbest function. The winning model was selected based on the R-squared statistic and the mean squared error. The winning model was the stepwise model. There were a number of variables in the dataset that added and subtracted from the number of possible wins. Ultimately the variables that played the largest role in obtaining a win were getting to 1st or 3rd base, getting a home run or walk, pitching a strikeout and not getting errors in the field.