ACSC 505 Group Project

Nick Johnson

Maryville University

12/3/2015

**Baseball Team Win Factors**

**Abstract**

In this project we will explore the contributing factors of baseball team wins. We looked at the baseball team wins in the MLB for the 2014 regular season. There are an abundant amount of baseball statistics we could look at but we limited it to six.

**Introduction and Background**

We think there are six factors that will help predict team wins for the regular season. The six factors and definitions as followed*: team batting average* which is all team hits divided by team at bats, *on-base percentage* is the total of hits, walks, and hit by pitcher divided by the total of at bats, walks, hit by pitcher, and sacrifice flies, *slugging percentage* is the total bases acquired by each at bat divided by the total at bats, *successful stolen base percentage* is stolen bases divided by the total of stolen bases and caught stealing bases, *strikeouts per game average* is the amount of strikeouts against the opposing team divided by the total number of at bats for the opposing team, *fielding percentage* is a measure that reflects the defensive teams handling of the baseball which is the total of putouts and assists divided by the total of putouts, assists, and errors. Our purpose of this report is to better predict team win rates for the following year.

**Method**

We collected data from MLB.com for defensive and offensive statistics. We use simple linear regression to analyze the data. We use SPSS software to generate the statistical information needed for this research project.

**Offensive Data**



**Defensive Data**



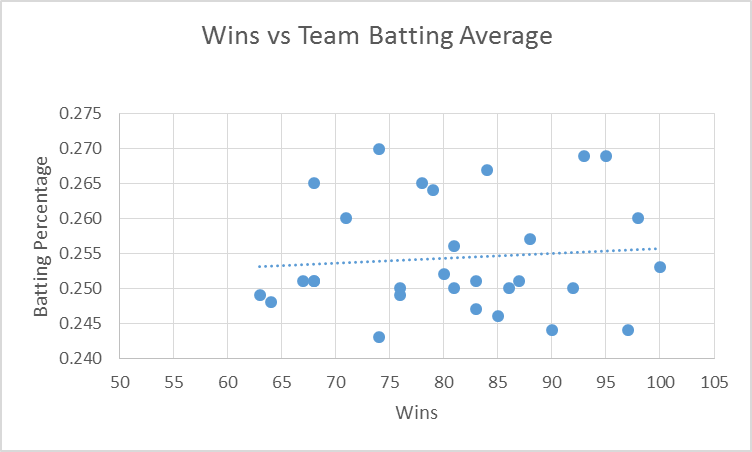
**Data Analysis**

1. Write down the multiple linear regression model with all assumptions.

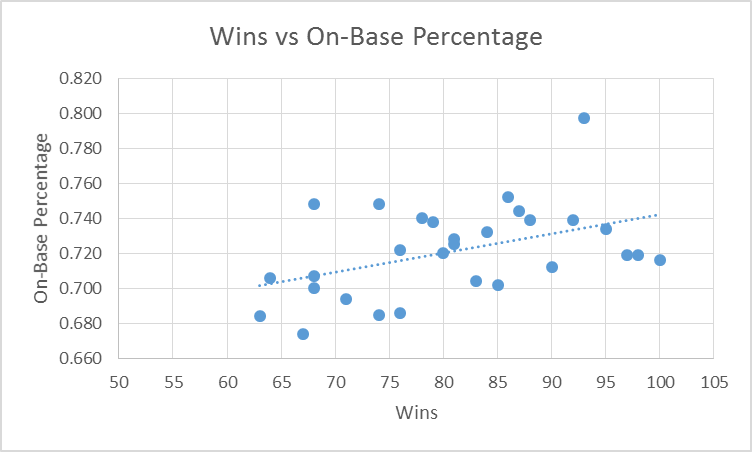
y = β1 + β2(Team Batting Average)x2 + β3(On-Base Percentage)x3 + β4(Slugging Percentage)x4 + β5(Successful Stolen Bases)x5 + β6(Strikeouts Per Game)x6 + β7(Fielding Percentage)x7 + ε

Ɛ~N (0, σ2) Ɛ follows independent identical distribution.

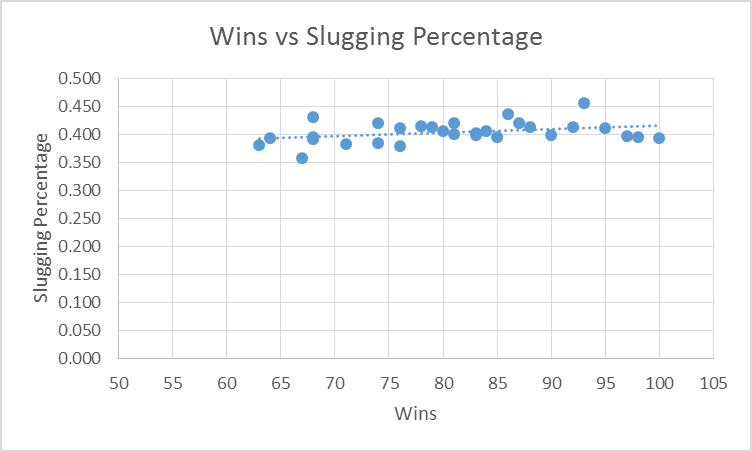
1. Obtain a scatter plot of each of the above and graph the respective estimated regression line on each plot. Comment on your plots.



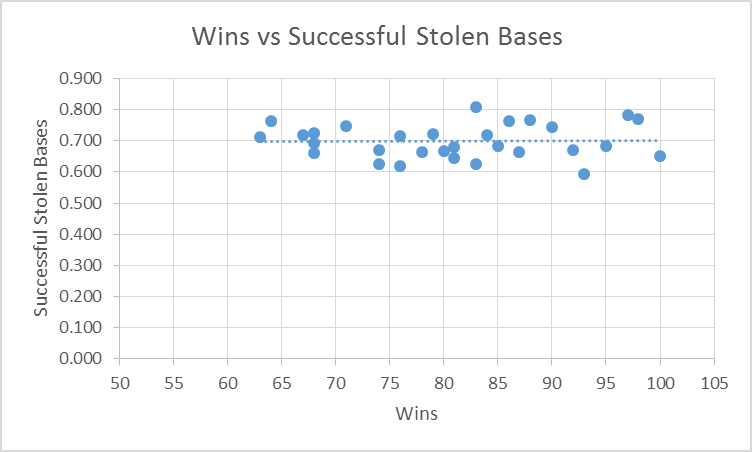
Comments: Even with outliers, the higher batting average should result in more team wins but the variance is high.



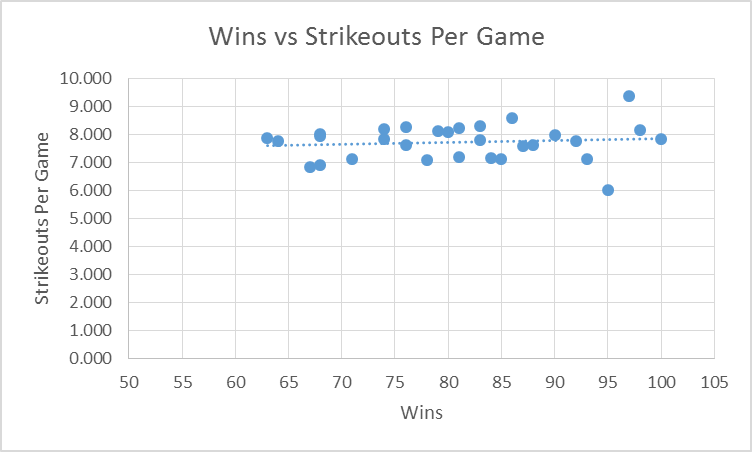
Comments: The variance is still a little high but the higher on base percentage of team’s results in a higher win rate.



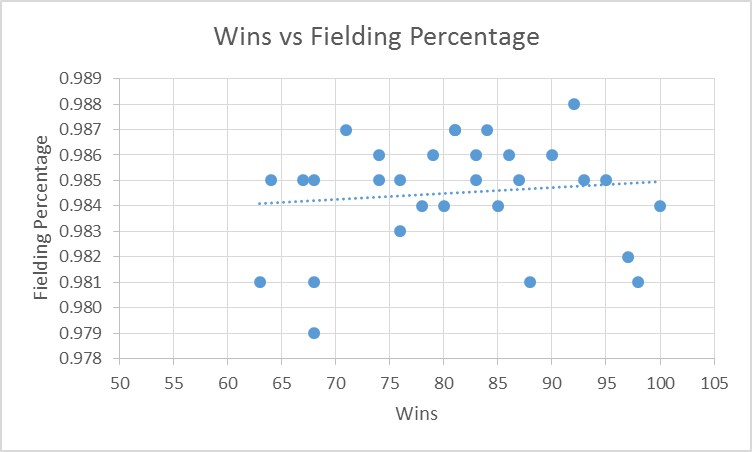
Comments: This has very little variance between slugging percentage and teams wins. But there is a slight increase in team wins in relation to team slugging percentage.



Comments: There is very little to no change in the successful stolen bases to team wins.



Comments: There is a slight increase in team wins with the increase in strikeouts per game.



Comments: A slight increase in team wins with an increase in fielding percentage. There are many outliers and this has a high variance.

1. Calculate, adjusted, Standard Error of the Estimate?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summary** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .671a | .451 | .308 | 8.69858 |

Predictors: (Constant), Field\_Per, SO\_PerGame, OnBase\_Per, Success\_SB, Batting\_Ave, Slug\_Per

1. Identify MSE for each of the three regression lines. Which independent variable leads to the smallest variability around the fitted regression line?

Y=β1+β2(Team Batting Average)X2 MSE=112.303

Y=β1+β3(On-Base Percentage)X3 MSE=90.031

Y=β1+β4(Slugging Percentage)X4 MSE=99.200

Y=β1+β5(Successful Stolen Bases)X5 MSE=113.166

Y=β1+β6(Strike-outs Per Game)X6 MSE=111.803

Y=β1+β7(Fielding Percentage)X7 MSE=111.756

On-Base Percentage leads to the smallest variability around the fitted regression line.

1. Using  as a criterion, which independent variable accounts for the largest reduction in variability?

Team Batting Average = .008

On-Base Percentage = .205

Slugging Percentage = .123

Successful Stolen Bases = .0001

Strike-outs Per Game  = .012

Fielding Percentage  = .013

Successful Stolen Bases has the lowest  value which is the largest reduction in variability.

1. Set up the ANOVA table based on lowest MSE.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 648.088 | 1 | 648.088 | 7.198 | .012b |
| Residual | 2520.878 | 28 | 90.031 |  |  |
| Total | 3168.967 | 29 |  |  |  |

|  |
| --- |
| a. Dependent Variable: Wins |
| b. Predictors: (Constant), OnBase\_Per |

1. Conduct an F test for condo of whether or not β3=0. Control the  risk at 0.01. State the alternatives, decision rule, and conclusion.

H0: β3 = 0 VS Ha:β3 ≠ 0

α = 0.01

F = 7.198

F0.01(1,28) = 7.64

F < 7.64, so Fail to Reject H0.

X2 isn’t significant in regression model at 1% level of significance.

1. Obtain a confidence interval level of 90% for β3.

x-bar = .721

F0.05(1,28) = 4.20

S.D. = 0.109545

n = 30

C.I. = (.637, .805)

1. Set up ANOVA table with decomposition of SSR.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 1428.664 | 6 | 238.111 | 3.147 | .021b |
| Residual | 1740.303 | 23 | 75.665 |  |  |
| Total | 3168.967 | 29 |  |  |  |

|  |
| --- |
| a. Dependent Variable: Wins |
| b. Predictors: (Constant), Field\_Per, SO\_PerGame, OnBase\_Per, Success\_SB, Batting\_Ave, Slug\_Per |

1. What are the values of β for the multiple linear regression model?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | -242.614 | 756.477 |  | -.321 | .751 |
| Batting\_Ave | -699.458 | 322.085 | -.546 | -2.172 | .040 |
| OnBase\_Per | 937.998 | 261.218 | 2.287 | 3.591 | .002 |
| Slug\_Per | -836.265 | 316.927 | -1.520 | -2.639 | .015 |
| Success\_SB | 38.779 | 33.021 | .200 | 1.174 | .252 |
| SO\_PerGame | -.982 | 3.259 | -.060 | -.301 | .766 |
| Field\_Per | 146.059 | 770.175 | .031 | .190 | .851 |

|  |
| --- |
| a. Dependent Variable: Wins |

1. Write down the linear regression model with all assumptions.

Y(wins) = -242.614 – 699.458X2(Team Batting Average) + 937.998X3(On Base Percentage) – 836.265X4(Slugging Percentage) + 38.779X5(Successful Stolen Bases) – 0.982X6(Strike outs Per Game) + 146.059X7(Fielding Percentage)

1. Plot the Y observations against the fitted values. Does the response function provide a good fit?



Overall the fitted values work pretty well. There are a few teams that are drastically different from the actual Y observations. The regular season consists of 162 games; having a few games plus or minus from the test Y observation compared to actual Y observation is really impressive. The few outliers are to be expected but if more variables are introduced perhaps the test Y observation could be more accurate.

**Results and Conclusions**

We chose a few offensive and defensive variables to analyze in SPSS. We didn’t want to make the variables one sided when looking at team wins, we wanted a variety of statistics that seemed logical from a personal stand point. According to the output we have from SPSS, we can predict the total wins a baseball team will get in a single year based on statistics acquired from MLB.com. The tested Y observations were close to actual Y observations with the exception of a few. We believe that the model could be very accurate with the addition of more independent variables.

**Data Source**

[www.MLB.com](http://www.MLB.com)

**Resources**

<https://en.wikipedia.org/wiki/Major_League_Baseball_schedule>

<http://www.baseballstatmanager.com/stat-list.aspx>

<http://baseballtips.com/glossary/index.html/>