MLB Team Success: Offense vs. Defense

MLB Data: 2012-2018

Andrew Argeros, Ryan Brauer, Jake Dujmovic

HAMLINE UNIVERSITY

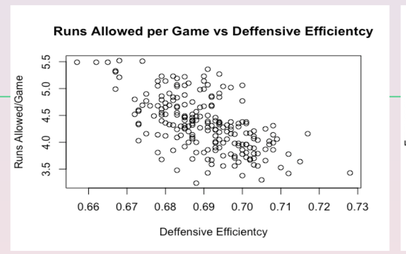
Abstract:

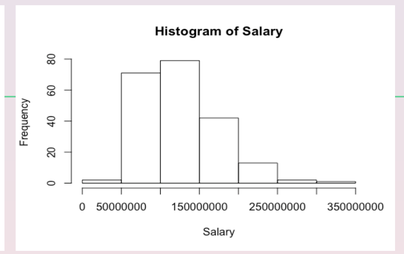
Data use in Major League Baseball (MLB) is crucial. There are more than 85 statistics being tracked over the course of 162 games, across 30 teams. The MLB captures data on every pitch of the 7 month season. In 2003 the idea of sabermetrics was popularized and later made famous by the movie 'Moneyball' in 2011. Since the early 2000's, data analysis in Major League Baseball has evolved rapidly and grown in popularity because teams can quickly find players that are valuable to them based on their statistics. Because teams are focusing so much on statistics in order to build a successful team, we examined: **which MLB statistics have the greatest impact on teams success in terms of win-loss percentage?**

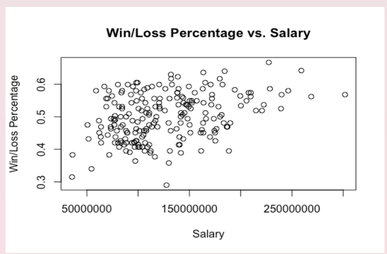
In 2005, Adam Houser researched this very question and found that OBP (on-base-percentage) and WHIP (walks and hits per innings pitched) were the most important and contribute most to an MLB teams success. We update Houser's analysis by looking at Kaggle data from the 2012-2018 MLB seasons, and analyzed with probit and logit models, regressing offensive and defensive MLB team statistics against team Win-Loss Percentage.

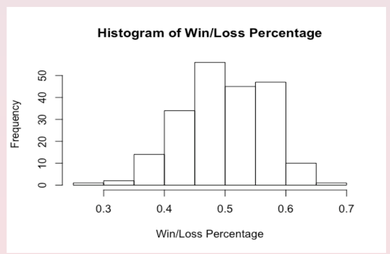
Today, we believe MLB teams and players have been focusing more on offense and power numbers. **We hypothesize that offensive statistics are most important in determining a team success.** However, we found that offensive statistics were not statistically different compared to defensive and pitching statistics when determining a team's win-loss percentage. Offensive statistics do not contribute significantly more to team success as we, and modern day baseball, had thought.

Descriptive Statistics:









Next Steps and Limitations:

The data set we chose to use had several of the basic statistics used in baseball, however there are several more advanced statistics that are not as convenient to get into a data set. Further testing of our hypothesis would incorporate more advanced sabermetrics such as fan graph pitch values and weighted runs created plus.

Sample size may be a limitation in our analysis. We have seven years of data however the modern era of baseball goes back to the 1930's. Having data for the last 90 years could give us a more accurate model.

Another direction further research could go is towards teams success and their salary. Are teams spending there money efficiently and on the right people and statistics? Where does a majority of the teams money go, offence or defense/pitching, and how successful are they?

Background and Hypothesis:

Data is everywhere in sports. Analytics is used to help teams make decisions and be more efficient in winning games. We see this in baseball more than any other sport right now. So what helps baseball teams win games the most? In 2005, Adam Houser wrote "Which Baseball Statistic Is the Most Important When Determining Team Success?" He found that WHIP and On Base Percentage were the two key stats that determine success. One is a pitching metric, while the other is an offensive statistic. However, the game has changed since 2005. Baseball has seen an increased focus on offensive statistics and offensive production. **As a continuation of Houser's research we propose that in today's game, offensive statistics are more important in determining a team's success (Win/Loss ratio) rather than pitching and defensive statistics.**

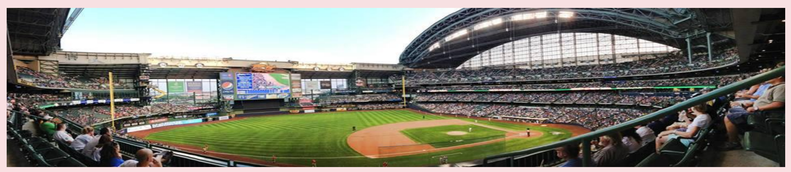
Methods and Assumptions:

To test our hypothesis we will be using a few different methods. We will begin by partitioning the data into validation and test data. 2012-2017 is test data and 2018 is validation data. We ran logit and probit models to analyze our data. For each type of model we ran a defense only, offense only and all stats model. An F-test was used to determine the significance, if any, from the models we ran. As we were testing the various probit and logit models, we used J.T. Clark's MLB Regression Analysis article and his analysis process to refine our model to come up with our most accurate model, the logit defensive model.

- There was no evidence of heteroskedasticity in our model when plotting, so there is no need for robust standard errors.
- Some multicollinearity is expected because most of the statistics in baseball are related to one another.
- There are so many different statistics in baseball that it is almost impossible to avoid omitted variable bias without overloading the models with variables. We are aware that there are most likely some variable we left out.

Importance:

Baseball is the most data driven sport today. We have seen a massive change in the way that data is used and the value of this data. MLB publishes 85 different individual statistics, and most of these can be combined for team statistics. Because it is such a long season, teams and managers need to rely on these stats to value a individual player or team. This allows teams to then allocate their money on the players they feel have the most value. We tested to see which statistics truly offer the most value.



References:

- Pelcastre, O. (2018). MLB Team Statistics. Retrieved from https://www.kaggle.com/omipelcastre/mlb-team-statistics.
- Houser, A. (2005). Which Baseball Statistic Is the Most Important When Determining Team Success? Retrieved from https://pdfs.semanticscholar.org/f1b1/aaafdbb70c03681b4738dc696cf9355dce69.pdf.
- Clark, J. T. (2016, May 5). Regression Analysis of Success in Major League Baseball. Retrieved from https://scholarcommons.sc.edu/cgi/viewcontent.cgi?article=1102&context=senior_theses.
- Castrovince , A. (2018, May 7). Poll of 70 big leaguers reveals what they value to gauge performance. Retrieved from https://www.mlb.com/news/mlb-players-vote-for-stats-they-value-most-c274986480.

The Data:

The data set we chose is from Kaggle and contains MLB team data from 2012-2018 with 28 variables for 210 observations. The data includes a mix of offensive and also pitching/defensive statistics as well as the total payroll for each team. The dependent variable we focused on was team Win/Loss Percentage. The data includes these statistics/variables:

Offensive Statistics: R/G, R, H, RBI, SB, SO, BA, OBP, SLG, GDP, LOB, the log of Salary (to minimize variance in the variable) , WAR, and we then added an OPS (SLG + OBP) variable, and a MVP dummy variable which allows us to recognize teams who had MVP's in their respective league that year.

Defensive Statistics: RA/G, DefEff, E, DP, ERA, Sho, H, ER, HR, BB, SO, WAR, salary, in addition to a CY Young variable so we can recognize teams with the best pitchers in their respective league.

Variables	Descriptions	Mean	Median	Min	Max
OPB	**On-base percentage**	.3186	.3185	.29	.349
BA	**Batting Average**	.253	.2525	.226	.283
ERA	**Earned run average**	4.036	4.000	2.94	5.36
DefEff	**Defensive efficiency**	.6901	.6901	.657	.728

We used WAR and the log of team salary as both offensive and defensive variables to effectively control for these variables.

Note: WAR is calculated using the FanGraphs formula.

Results:

Regression Results from Defensive Only Logit Model:

	Estimate Std.	Error	t value	Pr(>|t|)	
(Intercept)	-12.651	4.993	-2.534	0.0122 *	
DefEff	**7.835**	**4.594**	**1.706**	**0.0900 .**	**(Significant at 10%)**
ER	-.00036	.001533	-0.240	0.8105	
RA.G	-.34331	.2275	-1.509	0.1333	
tSho	. 00089	.003881	0.230	0.8186	
H	.00111	.0007748	1.443	0.1509	
HR	.00055	.001005	0.550	0.5830	
E	.0003	.001324	0.229	0.8192	
BB	-.00005	.0003594	-0.161	0.8721	
DoublePlay	**.00252**	**.0008680**	**2.911**	**0.0041 ****	**(Significant at 1%)**
SO	**.00083**	**.0003407**	**2.455**	**0.0151 ***	**(Significant at 5%)**
salary	-.000000001	.000000001	-1.517	0.1312	
MVP	**.12640**	**.04899**	**2.581**	**0.0107 ***	**(Significant at 5%)**
CYYOUNG	.0258	.05268	0.489	0.6257	
lsalary	**.3266**	**.1294**	**2.523**	**0.0126 ***	**(Significant at 5%)**

F-statistic All Stats	DF: All Stats	F-statistic Def-only	DF: Def-only
3.30	150	0.29	179

These results suggest that there is no econometric benefit to specializing a regression to just defensive statistics.

Accuracy:

The accuracy from both regressions is compared to the Pythagorean Expectation, a common method for predicting win-loss percentage. This method is considered the "naïve" method of estimating win-loss metrics.

$$PE = 1/(1+(Runs\ Allowed^2/Runs\ Scored^2))$$

All Statistics	Defensive Only	Pythagorean Expectation
69.8%	77.8%	95.8%