Regression Analysis of Team Success in Baseball

The Impact of Salary Expenditures

on Measures of Team Success

Austin Carnathan

Sherwin Vassigh

SJSU, Economics
Spring 2017
Project Advisor: Dr. Liu

Table of Contents

Introduction 2

Glossary of Terms 4

Methods and Analysis 5

Data Summary Statistics: 6

Discussion of Regression Results 7

Conclusion 9

Tables and Figures 11

Appendix of Regressions 14

Works Cited 15

I. Introduction

The statistical dissection of the sport of baseball is as enjoyable to some as the sport itself is to fans. That stems from the fact that baseball, more than any other sport perhaps, is a numbers game. As much as luck is attributed to success in any sport as is talent, baseball particularly lends itself to enthused statisticians and baseball junkies alike to weed luck out and find the root of success, with the overarching goal of replication and repetition. Therefore it comes as no surprise that there is an abundance of empirical research done in the field. We, however, aim to clarify the relationship between certain individual and team statistics with the continuing success of a franchise, especially as it pertains to its expenditures on player wages. With the field being so heavily saturated with brilliant analysis, we have chosen this particular dual-pronged approach to our analysis. We aim to justify the massively inflated payroll budgets of certain teams -- historically, the NY Yankees and the LA Dodgers -- by comparing team's payroll figure to our benchmark of success.

Baseball statistics offer a significant amount of insight to individual, as well as team performance. In a saturated market such as the MLB and its free-agency, information is key for any franchise; having any advantage over opponent scouts may be a decisive factor in assembling a potent roster within a certain budget and making the playoffs. Information that is easily acquirable, presentable, and perhaps extrapolated can serve as the cornerstone of a team's strategy and management and lead the way to dynastic franchise success. We, therefore, aim to use this avenue of research to identify which commonly available baseball metric has bearing on team

success, and to further dissect whether team salary expenditures have any influence on said success metrics.

This subject is quite interesting and relevant to our research team, both as undergraduate Economics students and as avid sports fans. With the more modern and well-documented advent of sabermetrics, sports economics has risen to the spotlight. The significance of statistics is best displayed in the MLB free agency, where all parties involved with a player's contract negotiations point to their individual stats. As BleacherReport correspondent Mark Hauser describes, "in baseball, more than any other team sport, stats give you an accurate picture of a player's performance because of the limited interaction between its players, teamwork is less important, hence, there are fewer intangibles to take into consideration when evaluating a player's abilities." Despite being a team sport, individual statistics give you a very clear picture of a player's production, serving as crucial information to evaluate during contract talks.

To that end, baseball has a multitude of offensive and defensive individual statistics being recorded constantly. From the plethora of batting numbers, to the fielding, pitching and baserunning numbers, baseball is full of seemingly meaningless percentages and averages. We aim to determine which of these statistics have the biggest impact on overall team wins during the course of a season. By potentially identifying the critical statistics on team wins, we then aim to determine whether a more winning team necessarily carries a higher payroll by analyzing team payroll expenditures.

With baseball being such a numbers-oriented sport, there are inherently a variety of metrics to consider in searching for those that most impact team success. Based on preliminary research, there are a handful of critical offensive and defensive statistics that are generally used by pundits. On-base, fielding and slugging percentages, batting averages, and WHIP among others, are all examples of such fundamental statistics.

Our thesis is designed to initially test the accuracy of using certain baseball metrics to predict team success, with the end goal being to justify team salary levels using the same metrics. With baseball being just as much a team sport as an individual sport, these statistic will help gain information who may deserve certain salary levels. We hypothesise that teams with better team statistics will have higher payrolls. This research will shed light as to which of these statistics may be more effective in predicting success than others, while examining the tendencies of overspending teams in terms of wins.

II. Glossary of Terms

R	(+)	Runs scored by a team in a season
OBP	(+)	On-Base-Percentage [(hits+walks)/at bats]
SLG	(+)	Slugging Percentage [(hits+2B+{Home Runs*3})/AB]
RA/G	(-)	Runs allowed per game (denoted as R.1)
WHIP	(-)	(Walks + Hits) / Innings Pitched
DefEff	(-)	Percentage of balls in play converted into outs
Wins		Number of team wins in the 2015 regular season

III. Methods and Analysis

As previously outlined, there is a lot of baseball statistics from various credible bastball tracking sources; websites such as *baseball-reference.com* have plentiful statistics that are free on an easy-to-use website.

Based on our initial research, we have decided on a host of offensive- and defensive-based statistics to use for our analysis. Based mainly on the research papers of Adam Houser and Dennis Moy, we settled on the above outlined statistics to model our regressions. Our first step was to standardize all of our data so that we can make a more meaningful interpretation of our analysis results. By standardizing our variables, we free up the dataset from units and per cents and instead present our regression in terms of standard deviations away from the statistical variable's mean. This also allows for a simple, meaningful interpretation and comparison of results.

Our models were then built upon each other. We began by analyzing the effect of Runs Scored and Runs Allowed on Wins. Once studied, we noticed a positive, significant relationship between Runs Scored and a negative, significant relationship between Runs Allowed and Wins. The model used was as follows:

(1) $\text{Wins} = \beta_0 + \beta_1 \bullet R + \beta_2 \bullet R.1 + \mu_i$

From there we wanted to dissect Runs Scored and Runs Allowed to see what was a significant factor in determining them, as a derivative of offensive and defensive production. We found that OBP and SLG were positively and statistically significantly related to Runs Scored, and that WHIP and DefEff were negatively and statistically significantly related to Runs Allowed.

Therefore our offensive model was based upon the power of the players on each team as well as the ability of a batters to reach a base; it is as followed:

(2) Runs Scored = $\beta_0 + \beta_1 \bullet OBP + \beta_2 \bullet SLG + \mu_i$

As for the defensive model, opposite applies, as we targeted stats that track player's ability to reduce opponent scoring based on limiting runners off base and creating outs on playable balls:

(3) Runs Allowed = $\alpha_0 + \alpha_1 \bullet WHIP + \alpha_2 \bullet DefEff + \mu_i$

We were able to illustrate in our two introductory analyses that:

a. A team's wins over a course of a season are a function of the number of runs scored by its offense and the number of runs allowed by its defence;
b. The number of runs scored is a function of OBP and SLG; and
c. The number of runs allowed is a function of WHIP and DefEff;

Our team, then, aimed to conclude that winning is determined by those four variables -- OBP, SLG, WHIP, and DefEff. We were left with the final regression model:

(4) Wins = $\pi_0 + \pi_1 \bullet OBP + \pi_2 \bullet SLG + \pi_3 \bullet WHIP + \pi_4 \bullet DefEff + \mu_i$

IV. Data Summary Statistics:

	R	**RA**	**OBP**	**SLG**	**WHIP**	**DefEff**
Min	-16.294	-3.833	-15.000	-9.000	-1.871	-5.500
1st Qrt	-7.088	-0.958	-8.167	-2.591	-0.617	-1.375
Median	-2.647	0.563	-5.000	-1.188	0.589	0.125
Mean	-2.737	0.375	-3.889	-0.715	0.429	0.075
3rd Qrt	0.794	1.489	0.833	0.955	1.315	1.875
Max	21.118	4.375	11.667	8.818	3.952	4.750

V. Discussion of Regression Results

1. **Base Wins: Wins = $3.778*10^{-16}$ + 0.03976 * Runs Scored - 0.07964 * Runs Allowed**
2. **Offense: Runs Scored = 3.922e-15+1.986e-1 OBP - 7.577e-1 SLG**
3. **Defense: Runs Allowed = 1.946e-16 + 9.566e-01 WHIP+ 1.128e-01 DefEff**
4. **Final: Wins = $2.731*10^{-15}$ - 0.07507 * WHIP - 0.02188 * DefEff + 0.02664 * OBP + 0.01494 * SLG**

Our preliminary regression (1) aimed to explore both the relationship and significance between the number of wins over the course of a team's season and the number of runs scored and the runs allowed by the team. We knew the relationship was going to be important and hypothesised that it would be significant. Analysis of our regression results showed that both coefficients were extremely significant, which makes simple sense -- the more runs a team scores, and conversely the less runs a team allows, the more wins they may expect. We took note of the statistically significant variables, as well as the low SER and high R^2 values for this regression analysis, as our team predicted that these values will suffer as we worked to pinpoint exactly what baseball statistic contributed more to wins.

From here, we wanted to determine what exactly made the offense and defense tick, statistically speaking. Our research and analysis strongly hinted towards a set of four widely-available baseball statistics as being important to determining offensive production and defensive efficiency. Thus, we regressed the chosen variables to Runs Scored and Runs Allowed, respectively:

Regression (2) aimed to dissect offensive production, as we chose to include OBP and SLG as our regressors and Runs Scored as the dependent variable.

Regression (3) aimed to examine defensive efficiency, in which we chose to analyze WHIP and DefEff as our regressors and Runs Allowed as the dependent variable.

Analysis of those two models supported our choice, as the coefficients were statistically significant in both instances. As we further developed our model, we consolidated our regressions in to a more condensed, concise regression. Our primary regression, therefore, aims to interpret the effect of all of OBP, SLG, WHIP, and DefEff on team Wins over the course of a season.

As we showed Runs Scored and Runs Allowed to be statistically significant to Wins, these four baseball statistics are just as significant to the success of a team. Thus, Regression (4) shows that these four statistics are important to the continued success of a baseball team. This is because the more often a team can get on-base as a ratio of their at-bats (OBP) coupled with a metric measuring the power of a team's hitters (SLG) is significant to producing runs. Conversely, a low figure in our two sabermetric-based defensive statistics (WHIP and DefEff) contributes directly to less runs allowed by a team. Again, the number of wins expected is a function of Runs Scored and Runs Allowed; so, Regression (4) identifies the relationship between these offensive and defensive variables and the number of wins a team can anticipate, based on our data from the 2015 regular season.

In interpreting our regression results, it is clear to see the positive relationship between the offensive statistics and team wins as compared to the negative relationship between certain statistics and team wins. Standardizing our inputted data set means that we are measuring our output in terms of standard deviations away from the mean.

Our data is easy to present and expatiate on: any increase in our four variables will simply increase the expected amount of wins over the course of a team's season by the variable's coefficient, measured in standard deviations from the mean.

Looking at OBP, we can expect with an approximately 0.03 standard deviation increase in their OBP over the course of a season, that a team's success rate would increase by 1 standard deviation. Both defensive statistics, WHIP and DefEff are negative and small, but for teams wanting to be successful, the smaller the number the better. Even with these numbers being such small margins, over the course of a season, it could be the difference between having a team make playoffs and ending their season early.

VI. Conclusion

Major League Baseball team owners are in essence, running a business. To that end, they must be considered as profit-maximizers -- as all business owners inherently are. This means that the owners must field a team that can most easily sell-out ballparks, a team of all-star pitchers, big-time sluggers, and the like. Homeruns and strikeouts are the central objective of the owner in managing their team. Fans like to see pitchers shutting-out opposing batters, and they want to see their sluggers hit ball after ball out of the park.

The definition of success in a team sport, however, is very clear-cut. And it does not necessarily reflect the apparent definition of a profit-maximizing business. Owners must guide their team to win championships. In a team sport that is so focused and heavily dependent on individual performances, roster management and payroll

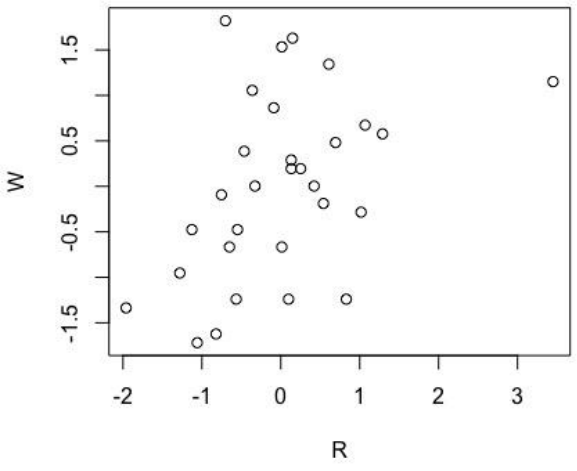
budgeting are of paramount importance. As the game of baseball is very simple, an owner must set up its team to utilize a finite amount of outs in each game in order to maximize the number of baserunners, all while limiting your opponent to the least amount of the same. Getting a player on base not only gets him closer to scoring a run, but it preserves an out for your team during his at-bat. It follows that the more players a team gets on-base during a game, the more likely they may be to win the game. The four variables we chose to research and analyze all measure this phenomenon. Coupled with our research, these stats may additionally highlight potential traits that players may have that would make them more prone to being successful by our benchmark. In implementing this, a team can expect more wins and therefore enjoy continued success.

In the process of our research, we stumbled upon many iterations of our study previously completed by other graduate- and undergraduate-level students. What we intended to achieve was different than anything we were able to preliminarily find, however our team had to come slightly short of our initial goals. We were not able to nail the analysis of multiple regression models we had come up with, namely to study the effect of team wins -- i.e. our four variables -- on team payroll, and to further dissect the payroll to determine if certain teams overpaid for the production they received over the course of the season.

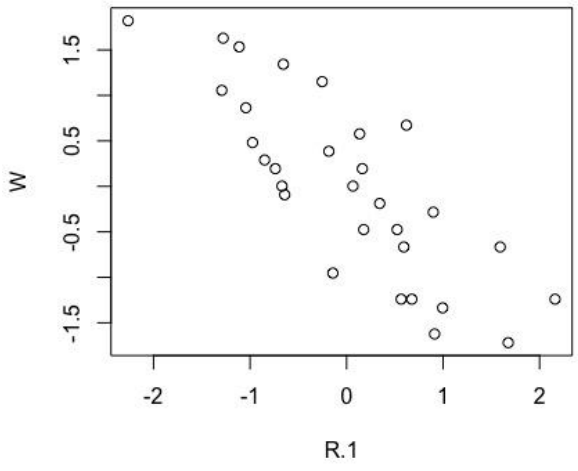
Our team had many high aspirations for this project, only limited by our time and current knowledge of powerful tools at our disposal. We hope to be able to further research the segment, exploring the effect of payroll budgeting and the weight assigned

to certain individual statistics when it is time to renew a player's contract, either correctly in the case of our four variables or incorrectly in the case of home runs or strikeouts. Additionally we would like to examine the long-term effect of continued team success that we have researched versus the current short-term, profit-maximizing tactic of hoarding and over-paying good pitchers and heavy hitters. This research may shed light on undervalued players that may even be in the free-agent market, who would benefit a team's greatly at a significantly lower price than the stereotypical overpaid all-star. Information such as this may help managers properly give weight to certain statistics, This analysis would have an overarching effect on long-term success, on the field and in the books.

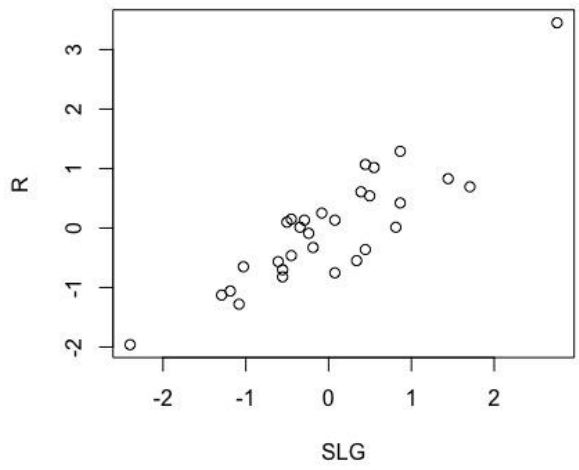
VII. Tables and Figures



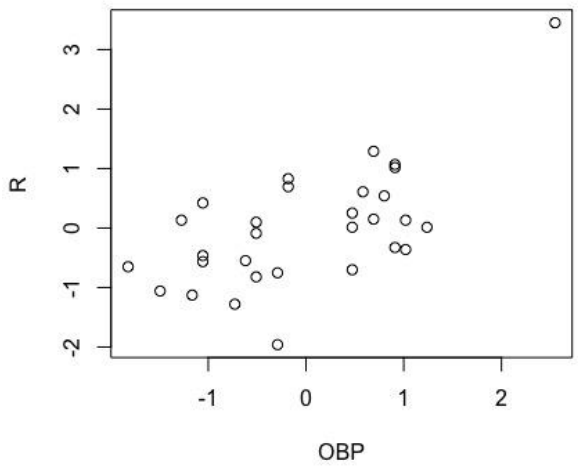
Wins vs. Runs Scored



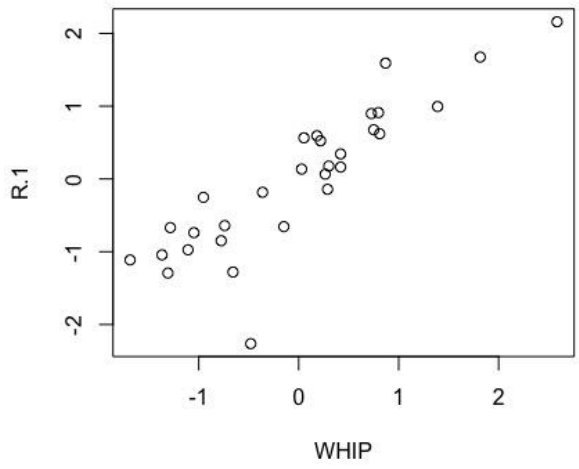
Wins vs. Runs Allowed



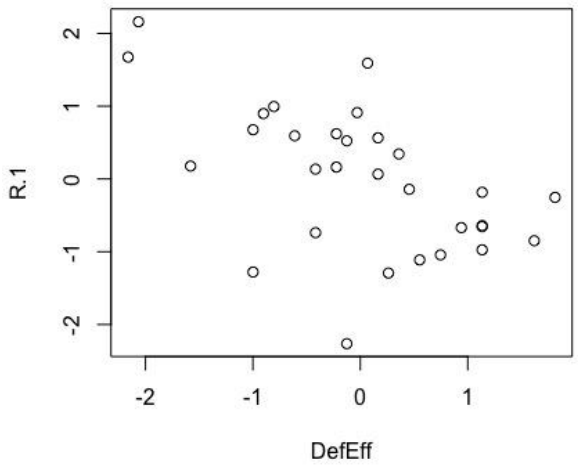
Runs Scored vs. SLG



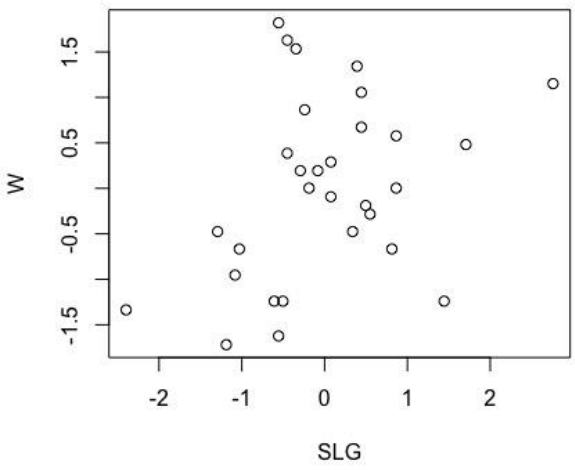
Runs Scored vs. OBP



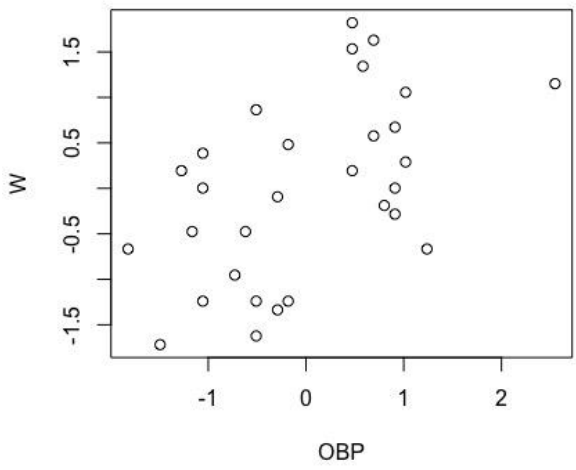
Runs Allowed vs. WHIP



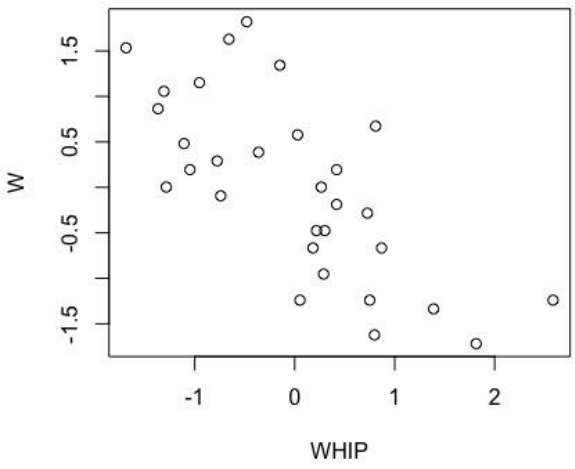
Runs Allowed vs. DefEff



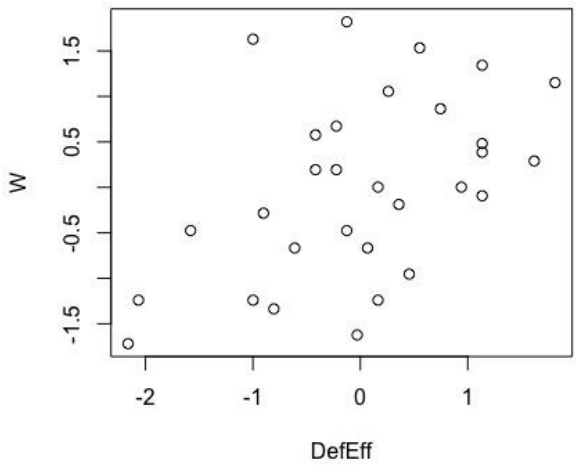
Wins vs. SLG



Wins vs. OBP



Wins vs. WHIP



Wins vs. DefEff

VIII. Appendix of Regressions

	Dependent Variable			
	(1) Wins	**(2) Runs Scored**	**(3) Runs Allowed**	**(4) Wins**
Regressor				
Runs	0.03.976(***) (0.008201)			
Runs Allowed	-0.07.964(***) (0.008201)			
OBP		0.01986(.) (0.0188)		0.02664 (0.01581)
SLG		0.07577(***) (0.0188)		0.01494 (0.01562)
WHIP			0.09566(***) (0.01275)	-0.07507(***) (0.01761)
Def.Eff.			0.01128 (0.01275)	-0.02188 (0.01841)
Intercept	$3.778*10^{-16}$ (0.008057)	$3.922*10^{-15}$ (0.008711)	$1.946*10^{-16}$ (0.008942)	$2.731*10^{-15}$ (0.01272)
Summary Statistics				
SER	0.4413	0.4771	0.4898	0.6583
R^2	0.8187	0.788	0.7767	0.6264
Adjusted R^2	0.8053	0.7723	0.7601	0.5667
n	30	30	30	30

IX. Works Cited

Hauser , Correspondent Feb 26, 2009, Mark. "Are Statistics Becoming Too Important in Sports?" Bleacher Report, 26 Feb. 2009. Web. 06 Dec. 2016. <http://bleacherreport.com/articles/130273-are-statistics-becoming-too-important-in-sports>.

Houser, Adam. "Which Baseball Statistic Is the Most Important When Determining Team Success?" *The Park Place Economist* 13 (n.d.): n. pag. The Park Place Economist. Web. 5 Dec. 2016. <https://www.iwu.edu/economics/PPE13/houser.pdf>.

Moy, Dennis. "REGRESSION PLANES TO IMPROVE THE PYTHAGOREAN PERCENTAGE." (2006): n. pag. University of California - Berkeley. Web. 5 Dec. 2016. <https://www.stat.berkeley.edu/~aldous/157/Old_Projects/moy.pdf>.