Analysis of the Economics of Baseball: Does More Money Get Better Results?

This analysis of Baseball Economics will mainly touch on two main interconnected questions. Does paying more money in salary get teams better results at the player level or at the team level? This study will be looking strictly at offensive production for its analysis of baseball economics for a couple reasons. One is that to win a baseball game a team needs to score runs, and while defense and pitching are key elements to winning at their best these aspects only keep a team from losing a game.

Secondly is that batting in Major League Baseball has a much higher failure rate than pitching and defense. Batting averages of .300 or higher are considered very successful which still gives a failure rate of 70%. If spending more money will actually improve the production teams would be wiser to spend their money on offense as the return on investment would be much higher than defense or pitching.

The data used in this study was obtained from Sean Lahman's website where he has compiled the statistics of every baseball season up to 2016. This study doesn't want to look at salaries from to far back as the effects of inflation will compromise any conclusions that could be drawn. So first the data was filtered to only include the latest five years of the data set which runs from 2012 to 2016.

Our data was split up into different categories such as batting, salary, fielding, etc. In order to get a master data frame to run calculations on these data frames had to be joined together. I used the batting statistics as my base data frame since this data frame contained the largest amount of pertinent information for each player

First thing done was to create a Batting Average column statistic in the Batting data frame since that was included in the original data. This was done by simply dividing a players total Hits by their total At Bats and rounded to three decimal places as is standard baseball practice. After that I merged the salary column from the Salary data frame with my Batting data frame to give each players production a dollar value. I did this on an inner join so if the player didn't have a matching salary to their player id and year id they would be dropped from the resulting data frame.

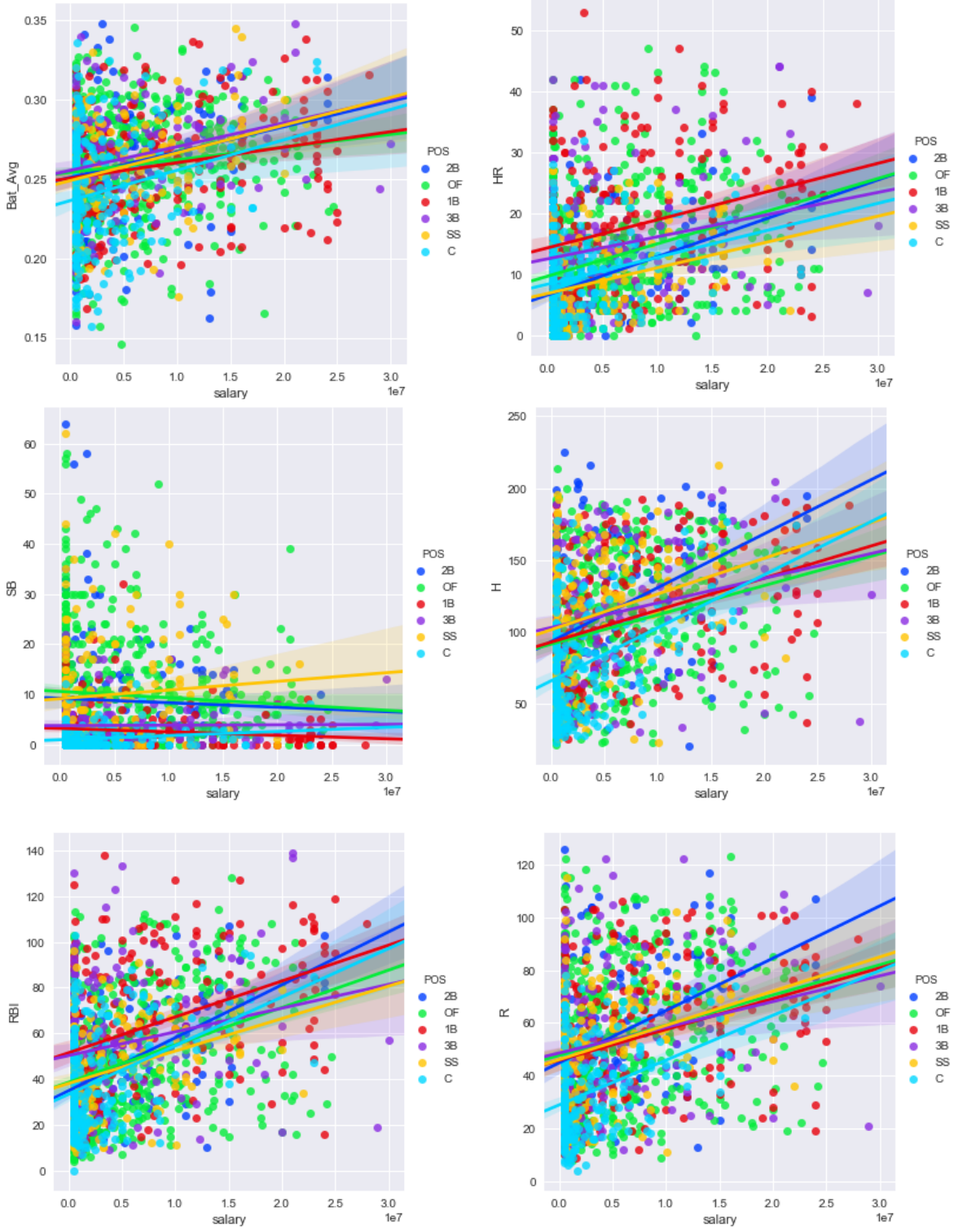
I felt this was better than including NaN values because while more production data would be nice their lack of salary would skew the averages by bringing down the average cost per metric. Next I wanted to add a fielding position to each player because that is not included in the original batting data frame as well. Doing so would be a little bit tricker though.

The original does have a fielding csv which shows each players position however it breaks it down by games played giving each player multiple positions per year. Simply merging this info with our master data frame would not work. But I didn't just want to pick a position at random out of the positions played and assign that to the player as that wouldn't be an accurate representation of their true position.

So the metric for determining a players position for each year was to find the max games played at one position for each player and return that position as a player's position for the year. Once this was done then I was able to merge the fielding data frame into my master data frame containing batting and salary info. My last change to the data was to add the player's names from the player master csv to my master data frame in order to obtain the player's full name if necessary.

In addition to this the player data was filtered to remove all pitchers as they are not paid on their batting contributions, and removed any player who did not achieve at least 130 At Bats a season. This was an arbitrary number chosen to match the number the MLB considers equal to the amount of playing time a player has to hit for his season to be considered his rookie year. Doing this removes players like pinch runners, and defensive replacements who are also compensated mainly on other factors other than hitting.

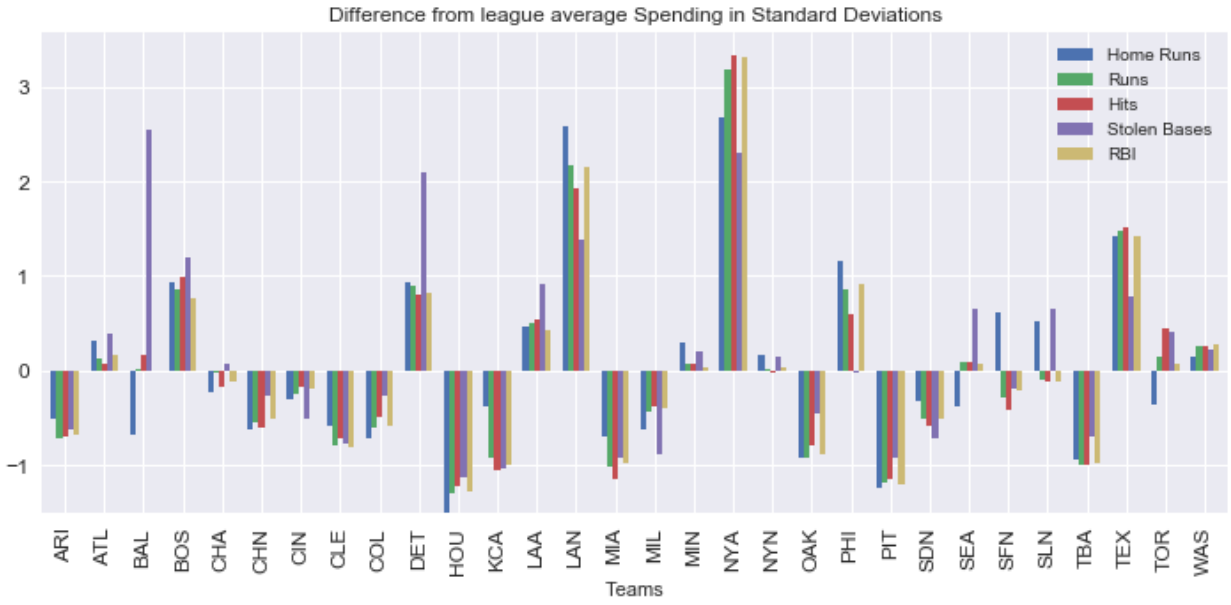
Now the master data frame was configured and held info on each players offensive production, salary, fielding position, and name. Using Seaborn's lmplot function I plotted each player's salary versus these offensive metrics: Home Runs (HR), Hits(H), Runs Batted In (RBI), Runs Scored (R), and Batting Average (Bat_Avg). Here are the resulting graphs:



Here we can see that there are positive correlations between salary and every offensive metric except for stolen bases. With stolen bases there is no relationship at all between salary paid and increased number of stolen bases. I have a suspicion that if one did a study on age and stolen bases there would be a much stronger correlation than salary, but that is for a different study.

Even though there are positive correlations for the other offensive metrics their slopes are generally not very strong. Given this I would say there is a weak relationship between salary paid and offensive production at the player level. But what about at the team level?

For looking at a team level I decided to correlate their win totals versus the amount of salary they spent standardized against the average salary spent. The reason I looked at wins as opposed to the similar statistics for the players is because a teams ultimate success is measured by wins and not their offensive productions metrics while a player can still be measured as a success by his offensive production even if his team loses. First I wanted to look at how much each team spent for each metric though just to get a feel for the data and see which teams overall where spending more money.

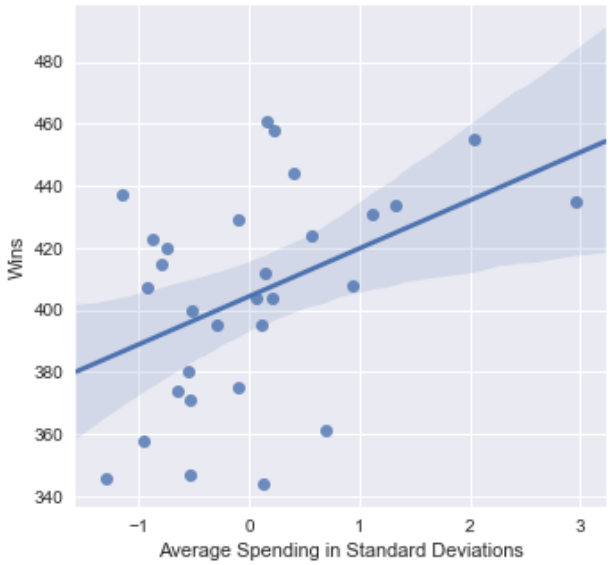


This graph shows that Boston, Texas, Detroit, L. A. Dodgers, and the New York Yankees all spent well above the average for each metric of player production, while Houston, Kansas City, Pittsburgh, and Tampa Bay spent well below. In order to achieve the stats in the graph I took my master data frame and grouped it by team in order to get the total of each offensive metric of the time span of the study. I then divided the total salary by total of each metric to get a cost per metric statistic. Basically I found out what each team was paying per Home Run, Hit, Run, etc. After that I then standardized each value and graphed them accordingly.

While the Yankees, Boston, and Dodgers are often contenders does spending more money correlate to actual wins? To find out I averaged each offensive productions standardized score to come up with an average score of spending above and below the mean and then plotted that against the wins. In order to do this though some more data manipulation had to be done.

I filtered the team records data by year and summed each teams wins to achieve the number of total wins and then appended that to my master data frame indexed by team that I used for the graph above. The results are seen below:

Spending on Offensive Metrics in Standard Deviations versus wins for years 2012-2016

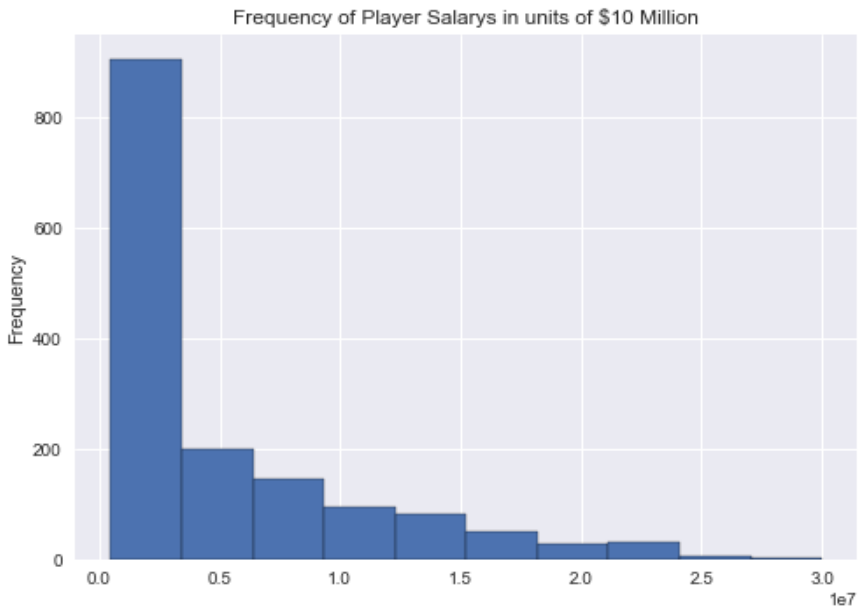


Again there is some positive correlation between spending more on offensive metrics and win totals. The graph data has a Pearson's R of .433 and a two tailed p value of .017. But there are still several factors outside of this graph that might further weaken the correlation. One is if you look at the top 5 and bottom 5 teams of the plot you'll see teams that under and over spent the mean in both selections.

teamID	W	STD_metrics_avg_by_team
SLN	461.0	0.166299
WAS	458.0	0.225697
LAN	455.0	2.042489
BAL	444.0	0.407795
PIT	437.0	-1.142529
PHI	361.0	0.698220
MIA	358.0	-0.956786
COL	347.0	-0.535232
HOU	346.0	-1.289481
MIN	344.0	0.130990

Here you can see that some of the biggest spenders from above such as the Yankees, and Boston are missing from the top five and that two teams that spent above the mean were some of the worst teams in baseball over this period. However the team with the most money spent the L.A. Dodgers achieved on average 91 wins a season a win total good enough to get a team to the playoffs most years.

Another reason to be wary about any correlation between spending and results at either a team or player level is the distribution of salary values in Major League Baseball as shown below.



The amount of salaries in the higher ranges to even take measures of are very small in comparison to smaller salaries which could lead to greater variances in our measurement of production for those high values. In effect the stars of baseball who make the larger salaries are outliers from the population and could be creating a positive correlation between spending and prouduction where there actually is none in the general population.

Ultimately the data shows a correlation between spending more money and achieving better results at a player level and at a team level in the MLB however there seems to be an effect of diminishing returns between the two variables esepcially at the team level as spending much more than league average defined by more than one standard deviation from mean on offense doesn't translate as much to wins as much as some of these MLB teams would like to think. My biggest take away from this is that teams should rely on more traditional methods of team development such as scouting and drafting to improve team performance rather than outbidding other teams for player talent on the open market.