Dollars for Doubles

An Econometric Analysis of MLB Salary On Offensive Performance

Sam Turner

Introduction

In 2019 Mike Trout set the MLB record for largest contract ever signed at $426 million. But with the league minimum at $700,000 in 2022, is it really fair to say that he performs at a level 609 times better than his "minimum wage counterparts? That is exactly the question this paper aims to address, to what extent is offensive performance, in the MLB, awarded in salary?

This paper is of particular interest to owners and baseball operations personnel in the MLB but is also of interest to the general public. Often when building new stadiums or renovating a current one, a city will financially support their team making the taxpayer's part owners in the franchise. Therefore, the general public should have an interest in how their money is being spent by these teams, since the teams have enough money for blockbuster deals but not enough for the stadium they want.

This paper uses a two-stage least-squares regression to first quantify a player's offensive performance in a given season. Then, the player's performance is compared to the real salary that they earned also in that season. The data used comes from the Sean Lahman relational MLB database. This database is free and available to the public. The data base has observations on a plethora

of statistics either directly or indirectly related to the MLB and its teams. This includes data on offensive outcomes for both teams and players from 1871 to 2021. There is also salary information that spans from 1985 to 2016. The intersection of the three of these data sets has over 15,000 observations.

Ultimately, this paper reaches the conclusion that there is a weak but positive relationship between offensive performance and real salary. This makes intuitive sense since offensive performance should be rewarded by the market, yet there are other factors that influence a player's salary. Defensive performance and fan opinion just to name a few. Further research to calculate the quantifiable aspects of a player's performance is needed to remove as much endogeneity and approach the true relationship of a player's performance to their salary.

Literature Review

This paper is not the first to consider how player performance is rewarded in the MLB. After the 1974 labor strike in the MLB Gerald W. Scully published his article *Pay and Performance in Major League Baseball* where he considers a similar topic. His goal is slightly different, focusing on measuring the economic loss of players due to the reserve clause in all MLB player's contracts. However, in order to reach this conclusion he must first calculate player's marginal revenue and uses simplifying assumptions of ratios to represent offensive and defensive performance, instead of OLS estimators of runs like this paper. His conclusions are similar to this paper's in that he also concludes that offensive performance is a poor sole explanation variable for player salary. However, his model improves greatly when he accounts for defensive achievements meaning that follow up papers on this paper's topic are likely to greatly help this paper's explanatory power.

More contemporary literature has, at best, come to mixed conclusions about the subject. Shahriar Hasan finds in *Can Money Buy Success ?: A Study of Team Payroll and Performance in the MLB,* there is strong and statistically significant evidence that a team's payroll effects their win percentage. His data went from 1995 to 2011. Yet, Stephen Hall found the opposite in his 2002 article, *Testing Causality Between Team Performance and Payroll: The Cases of Major League Baseball and English Soccer.* Hall finds that there is no causal link between payroll and team performance, measured in winning percentage. Interestingly, Hall does find correlation between payroll and performance. His data goes from 1980 to 2000. Finally, Jacob Andrew Loree takes the middle ground in his 2016 article, *Determinants of Baseball Success: An Econometrics Approach*. Loree finds that there is a statistically significant effect of salary on team success, measured by total runs and bases for a team in a season. However, Loree points out that the cost of buying these additional wins is too great for many small market teams to afford, making the effect basically nonexistent.

Thus, with three vastly different approaches and conclusions to consider, this paper aims at attempting to help settle this debate by considering more contemporary data.

Theory

Using microeconomic intuition, we would expect to find that there should be a logarithmic relationship between salary and performance. This is due to the law of diminishing returns. We would expect players to earn more at an exponential rate when they perform above average. This is because there is competition between all teams, both large and small market, for players who are slightly above average. This large amount of competition will drive their wages up. This exponential curve will eventually start to level off once the salary level reaches the point that only a few of

the largest market teams can afford the player. This decrease in competition is what will make the curve almost parallel to the x-axis at the greatest levels of performance.

We should also expect that the salary and performance should have a high correlation. Competition between teams for players should allow for the true value of the player to be reached under the efficient market hypothesis. Teams also have access to all the same information as one another with respect to player performance on the field. Therefore, this transparency of information should lead to teams having similar valuation of player's and their abilities.

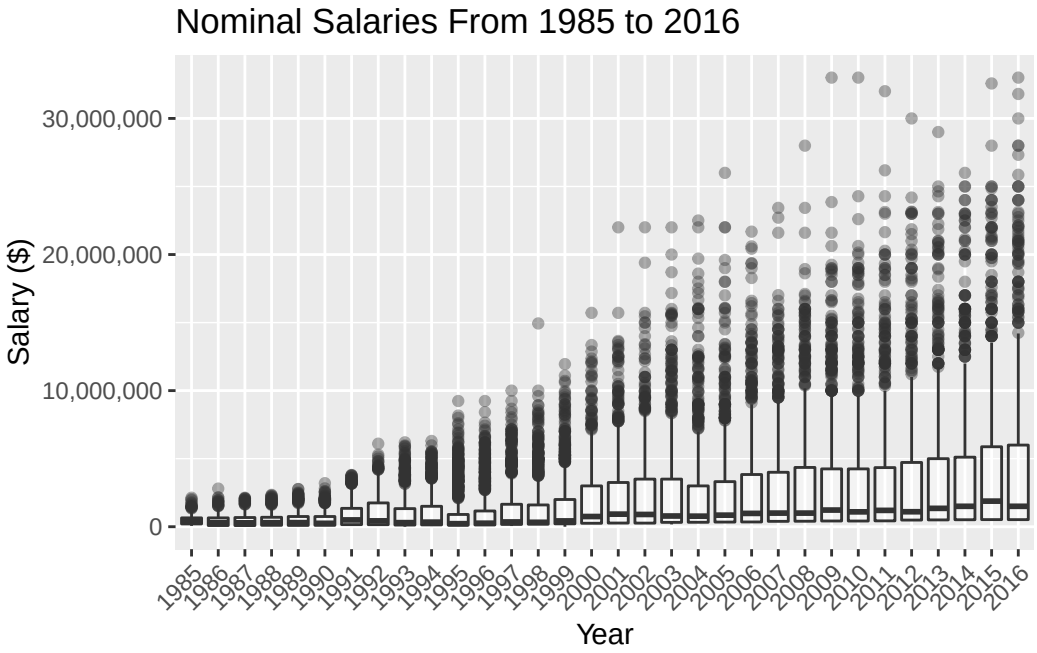
Data Description

There are two data bases that were used for this analysis. The first, for all the baseball statistics and observations, is the Sean Lahman database. And the second is the CPI index from the Federal Reserve Bank.

The Lahman database is extensive with over 30 datasets covering different aspects of the MLB from player performance to stadium details and more. It is a relational database meaning that every data set is related to another data set by a vector which has the same values for each observation. This feature is particularly helpful for this paper because it aims to use panel data of every player from 1985 to 2016 in an attempt to most accurately measure the player performance and salary. The main data sets that this paper uses are Salaries, Appearances, Teams, and Batting.

One of the few limitations of the Lahman database is that the Salaries data set only contains observations from 1985 to 2016. Therefore, this analysis will only use data observed from 1985 to 2016 across the other data sets. The salaries data set only has three variables of interest to this analysis, playerID and yearID to find a player's performance. And salary, which is the player's

nominal salary in a given year.



The Appearances data set was only used for filtering by position players. In the National League (NL) pitchers are required to hit to continue pitching the next inning. Including them in the analysis will create outliers since they are being rewarded for their pitching performance and not offensive performance. Thus, to eliminate them a player must have at least 5 appearances in a position other than pitching to be considered for the analysis.

The Teams data set includes almost 3000 observations on almost 50 variables. Since this paper is only on offensive performance, only a few are of interest. A couple variables must be calculated manually since they are not provided by the data set. These variables are OB, which is a statistic to cumulatively represent when a player reaches base without putting the ball in play. And X1B which is simply the total singles in a season, calculated by taking total hits and subtracting the count of every other hit outcome from that total.

Table 1: Teams Data Set Variables

Variable	Description
R	Total runs scored by a team in a season
X1B	Total singles hit by a team in a season
X2B	Total doubles hit by a team in a season
X3B	Total triples hit by a team in a season
HR	Total home runs hit by a team in a season
OB	Total times reached base without putting ball in play
SF	Total sacrifice flies hit by a team in a season

Table 2: Teams Summary Statistics

Variable	Min	Median	Mean	Max	Std. Dev.
R	466	729	729.99	1009	89.86
X1B	646	971	965.84	1186	77.97
X2B	159	276	275.29	376	34.16
X3B	6	30	30.97	61	9.08
HR	58	157	158.06	264	37.18
OB	350	571.5	574.71	841	77.51
SF	23	45	45.36	75	8.89

The Batting data set breaks down all of the offensive observations by player instead of by team as Teams did. This makes it almost identical to Teams other than it only considering only offensive

observations of players, hence the name Batting.

The only other data set used is the CPI index from the Federal Reserve Bank. This data set was accessed through the "quantmod" library. It describes change in CPI per month since 1947. However, since MLB salaries are only changed yearly at most, taking the mean CPI per year and using that as the CPI makes the most sense in this context.

Empirical Model

In baseball, the only way to win is to score runs. There are no ties so a team that does not score can only lose. Therefore, the best way to judge a player's performance is judging how many runs they can manufacture for their team. To succinctly measure a player's ability this paper creates a new statistic called Expected Runs Created (xRC) which is calculated using a multivariate fixed effects OLS regression which can be represented by the following equation.

$$xRC_{it} = \beta_0 + \beta_1 X1B_{it} + \beta_2 X2B_{it} + \beta_3 X3B_{it} + \beta_4 HR_{it} + \beta_5 OB_{it} + \beta_6 SF_{it} + \alpha_i + \tau_t + \epsilon_{it}$$

The first step to deriving xRC is finding the weights that each offensive outcome has on a team's runs. All offensive outcomes are not created equal, which makes intuitive sense. A single only "creates" a run when there is a player on second or third. While a home run guarantees a run and has the possibility of "creating" up to three more runs depending on how many players were on base when the ball was hit. This difference in value towards creating runs is represented in the equation as the β values.

A multivariate regression gives a decent estimate of the weights and removes some omitted variable bias from xRC as opposed to only considering hits. However, it violates one of the four major assumptions of the ordinary least squares (OLS) regression. Which is that there is no autocorrelation across observations. Since players do not resign to a new team every year, there is definitely correlation across team's performance from year to year since a team's performance is ultimately its player's performance. This means that controlling across time is needed at the least. This is represented in the equation by the subscript t's and the additional error term τ_t. There are also many factors in each team, such as team chemistry or culture, which are correlated with player performance and thus correlated with team performance. Therefore, controlling for team is also very important to remove endogeneity. Controlling for team is represented by the subscript i's and by the additional error term α_i.

Calculating for xRC is only the first step in determining if players are compensated for their offensive performance. The next step to answering the research question would be to run a bivariate OLS regression which can be represented by the following equation.

$$RealSalary = \gamma_0 + \gamma_1 xRC + \nu$$

In this equation γ_0 represents what would be expected of a player if they contributed no runs which should be league minimum salary. γ_1 would represent how much more money they could expect to earn for each additional run they produce. There is a lot of endogeneity in this $RealSalary$ regression especially from omitted variable bias. There are many measurable and unmeasurable variables left in the error term. For example, a measurable factor left in the error term would be defensive performance. Defense definitely plays a factor in a player's salary since a team might not

have to score as many runs in a game if they have a solid defender to prevent runners from reaching base. An unmeasurable variable in the error term that is causing endogeneity would be public opinion of a player. Even if a player is quite skilled they may have to sign for less than their xRC would estimate because of public outcry. Due to this endogeneity and omitted variable bias it is expected that there is a somewhat weak but positive relationship between xRC and $RealSalary$.

Thus, $RealSalary$ is not expected to be the final say on this question since it contains so much endogeneity. Rather, it aims merely at attempting to explain how much offensive contributions by players are rewarded in their salary. Further research should be conducted to estimate how defensive and pitching performance factor into player's salaries as well.

Results and Implications

```
[1] "CPIAUCSL"
```

Surprisingly, it can be hard to tell which model is best going only off of the outputs of the models. They are all tell a very similar story. However, using econometric intuition, the two way fixed effects model should be giving the most unbiased estimators of offensive outcomes because it controls for the most endogeneity. Therefore, we can substitute the β's from before and get the following equation,

$$\widehat{xRC} = .47X1B + .69X2B + 1.22X3B + 1.41HR + .32OB + .85SF$$

The coefficients in xRC represent how many runs are created per outcome. As expected, the weight for home runs is both higher than the weight for singles and is greater than 1. All weights

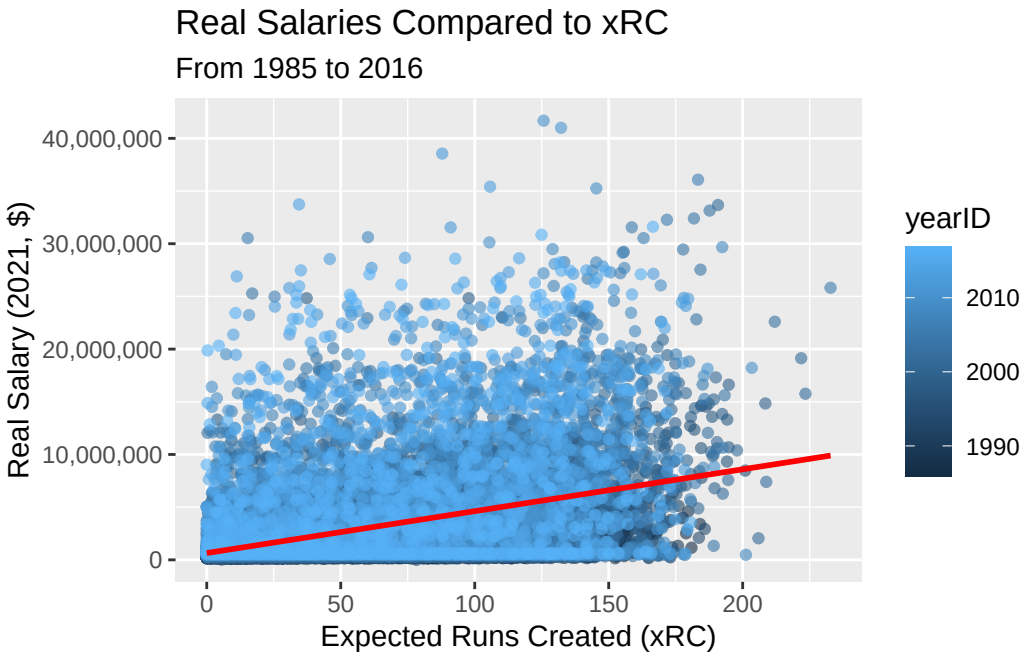
	Baseline	Multivariate	Control Time	Control Team	2 Way Fixed Effects
Constant	-244.00^{***}	-204.65^{***}	-422.78^{***}	-156.56^{***}	-389.93^{***}
	(23.98)	(13.82)	(17.19)	(17.50)	(20.63)
H	0.68^{***}				
	(0.02)				
X1B		0.34^{***}	0.50^{***}	0.31^{***}	0.47^{***}
		(0.01)	(0.01)	(0.01)	(0.02)
X2B		0.46^{***}	0.67^{***}	0.48^{***}	0.69^{***}
		(0.03)	(0.03)	(0.04)	(0.04)
X3B		1.10^{***}	1.16^{***}	1.20^{***}	1.22^{***}
		(0.11)	(0.09)	(0.13)	(0.10)
HR		1.37^{***}	1.43^{***}	1.34^{***}	1.41^{***}
		(0.03)	(0.03)	(0.03)	(0.03)
OB		0.28^{***}	0.32^{***}	0.30^{***}	0.32^{***}
		(0.02)	(0.01)	(0.02)	(0.01)
SF		1.49^{***}	0.97^{***}	1.26^{***}	0.85^{***}
		(0.13)	(0.10)	(0.13)	(0.11)
N	918	918	918	918	918
Adj. R^2	0.64	0.89	0.94	0.90	0.94
SER	53.57	29.37	22.12	27.22	20.87

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

	Real Salary Reg
Intercept	643 803***
	(58 590)
xRC	39 739***
	(746)
N	15 112
Adj. R^2	0.16
SER	4 264 718.86

* p < 0.1, ** p < 0.05, *** p < 0.01

are very statistically significant and the adjusted R^2 is also quite high. xRC's adjusted R^2 explains about 95% of the variation in runs scored which means that is should by a very good estimator for player's runs created.



The model measuring Real Salary using xRC as a predictor finds that for every one run increase in output, a player can expect to be compensated an extra $40,000. In addition, the intercept of this bivariate regression makes intuitive sense too, since the minimum amount a player can make changed from 1985 to 2016 floated around $500,000 to $600,000 when adjusting for inflation. Substituting these values, into the $RealSalary$ equation, it becomes,

$$\widehat{RealSalary} = (\$643803) + (\$39739)xRC$$

While this model does find both γ_0 and γ_1 statistically significant, it should be obvious that these coefficients are not practically significant. The model can only use xRC to describe about 15% of the variation of salary as it related to xRC. In addition, the standard error of the regression (SER) is about \$4,250,000. This means that this regression is about \$4 million off on average for each data point. As discussed before, this is likely due to endogeneity created from omitted variable bias. Finding a reliable way to measure defensive performance would greatly help this model since that is the other main contributor to salary. Due to all of this endogeneity, it is likely that the true effect of runs created on salary is likely lower than what is estimate.

The fact that a model using runs created does not explain salary well is both a good and bad thing. It is bad because it means that owners are spending money for players who do not perform as well as they are being paid for. However, this can also be seen as a good thing because it means that there is opportunity for arbitrage. If some owners are willing to pay for players who are not worth their contracts, those owners need to let go of some other players who are a better deal to free up the cash. This provides opportunities for other clubs who have better analytics and information, to cheaply acquire real talent.

Bibliography