Dollars for Doubles

An Econometric Analysis of MLB Salary On Offensive Performance

Sam Turner

6 December 2022

Introduction

In 2019 Mike Trout set Major League Baseball's (MLB) record for largest contract ever signed, at $426 million over 12 years (Langs 2022) or $35,500,000 annually. But with the minimum league annual salary at $700,000 in 2022, is it really fair to say that he performs at a level 50 times better than his "minimum wage" counterparts? That is exactly the question this paper aims to address. In the MLB, to what extent is offensive performance rewarded in salary?

This paper uses a two-stage least-squares regression to find offensive performance's relationship to salary. To quantify a player's offensive performance in a given season, a multivariate two way fixed effects ordinary least squares regression is run to quantify how each offensive outcome impacts a team's ability to score a run. Then, those outcomes are rated accordingly. Players are evaluated by how often they make those outcomes happen. Then, the player's performance is compared to the real salary that they earned in that season. The real salary is calculated by adjusting the salary the player earned in that year to its 2021 value. 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 data sets containing information on teams, batting, and salaries has over 15,000 observations. And the data on inflation comes from the Federal Reserve.

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 such as 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* (Scully 1974) where he considers a similar topic. His goal was focusing on measuring the economic loss of players due to the reserve clause in player's contracts. However, in order to reach this conclusion he first needed to calculate player's marginal revenue. In his calculations he used simplifying assumptions of certain ratios to represent offensive and defensive performance, instead of an ordinary least squares (OLS) estimation of runs like this paper. His conclusions on offensive performance and pay 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 performance, giving hope for follow up research

for this paper.

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* (Hasan and Rivers 2011), 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, Szymanski, and Zimbalist 2002). 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 2016). 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 to help settle this debate by considering more contemporary data.

Theory

Using microeconomic intuition, we would expect to find that there should be a quasi-exponential relationship between salary and performance. This is because extremely talented players earn economic rent. When dealing with players at or near the average, teams have a lot of bargaining power because average players are easily replaceable. Thus, if a player improves a little when they are

around the average they will not see a large increase in their salary. However, teams lose this bargaining power when dealing with players who perform much higher than average because these players are not replaceable. These high performing players can demand salaries magnitudes higher than their average counterparts and teams will be forced to pay them because they would benefit from having that talent on their team. Incidentally, this explains why Mike Trout can justify his $426 million contract.

For bargaining situations not dealing with extremely talented players, we should expect a player's true value to be reached through competition. Since most players perform around average and all teams can compete to sign an average player, this competition should result in players being fairly compensated for their performance. This should be a simple supply and demand problem trying to reach an equilibrium point.

Data Description

There are two data bases 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.

The Lahman database is extensive with over 30 data sets 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 and team 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.

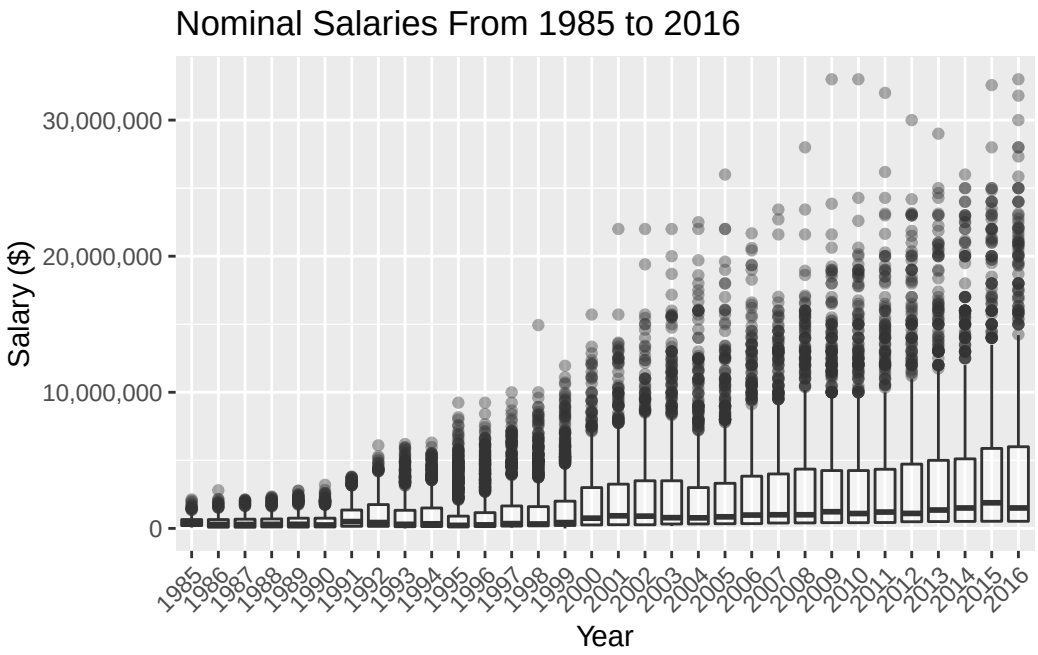

Figure 1: Nominal Salaries

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. These high paid pitchers would cluster with high salaries and low offensive production. 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 3,000 observations on almost 50 variables. Since this paper is only on offensive performance, only a few variables are of interest and are described in Table 1.

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. It is calculated by adding total observations of hit by pitches, walks, and intentional walks. And X1B which is simply the total singles in a season, calculated by taking total hits(H) and subtracting the count of doubles(X2B), triples(X3B), and home runs(HR).

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 (Walks[BB] + Intentional Walks[IBB] + Hit By Pitches[HBP])
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

Variable	Min	Median	Mean	Max	Std. Dev.
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 considers 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 (X1B) only "creates" a run when there is a player on second or third. While a home run (HR) guarantees a run will score 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 different β values. The ϵ in the equation is the error term from this multivariate regression and represents all other variables not included that influence xRC.

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, only using the multivariate equation violates one of the four major assumptions of an OLS regression. Which is that there is no autocorrelation across observations. Since players do not re-sign 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 the culmination of its player's performance. This means that controlling for teams is needed. This is represented in the equation by the subscript i's and the additional error term α_i. There are also factors that affect all teams equally that change from year to year. This can be something as simple as a rule change. The MLB has had many rules changes over its lifetime, including since 1985 when the data this paper considers begins. Thus, there is a need to include a control for time. This is represented in the regression equation with the subscript t's and the error term τ_t.

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 to compare player performance and pay 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, ν. 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. Players who are good at offense also tend to be good at defense. Players often practice to two together which would make them correlated and satisfy the requirements for omitted variable bias and thus create endogeneity. 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 the negative public opinion for the team signing them. If public opinion is low for a player they will also probably be heckled by the fans which can lead to a performance drop reflected in xRC. This situation would also satisfy the requirements for omitted variable bias. Due to this endogeneity and omitted variable bias it is expected that there is a somewhat weak but positive relationship between xRC and $RealSalary$.

Table 3: xRC Regressions Summary

	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

The $RealSalary$ regression 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 in Table 3. They 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 use the weights from the two way fixed effects model and substitute them into 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 predicted, the weight for home runs is both higher than the weight for singles and is greater than 1. All weights are very statistically significant and the adjusted R^2 is also quite high. xRC explains about 95% of the variation in runs scored which means that is should by a very good estimator for player's runs created.

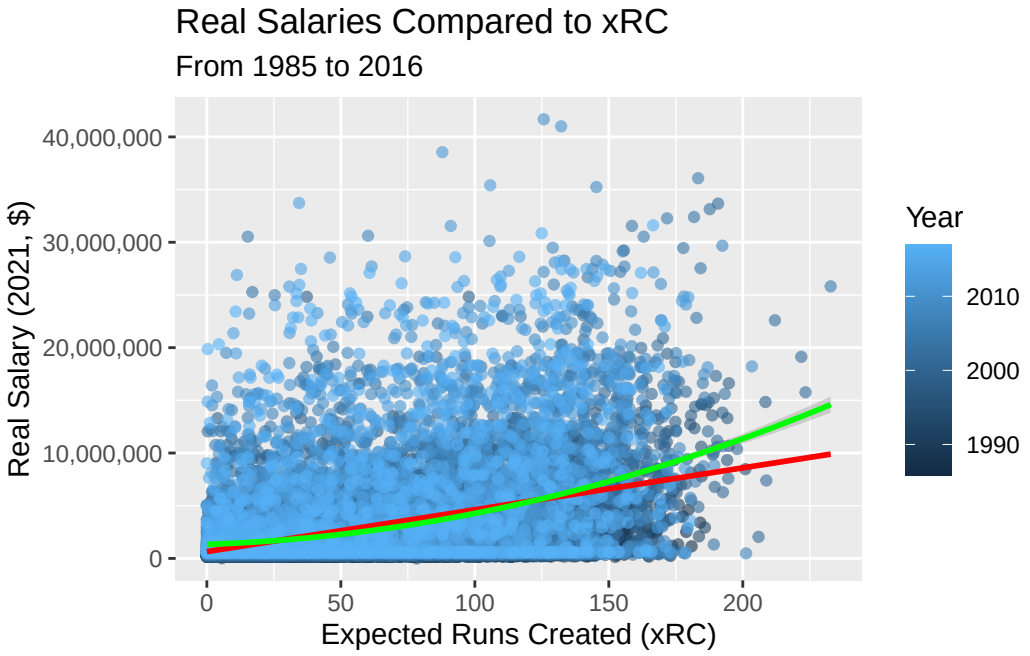

Figure 2: Real Salaries Regressions

The linear model (red line Figure 2) 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 (see

Table 4: RealSalary Regressions Summary

	Real Salary Reg	Real Salary Sqr. Reg
Intercept	643 803***	1 332 481***
	(58 590)	(79 229)
xRC	39 739***	8622***
	(746)	(2536)
xRC^2		208***
		(16)
N	15 112	15 112
Adj. R^2	0.16	0.17
SER	4 264 718.86	4 241 674.40

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

Table 4). In addition, the intercept of this bivariate regression makes intuitive sense too, since the minimum amount a player can make from 1985 to 2016 was around \$500,000 to \$600,000 when adjusting for inflation. Substituting these values, into the $RealSalary$ equation, it becomes,

$$\widehat{RealSalary} = (\$643,803) + (\$39,739)xRC$$

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 16% 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 line 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 lower than what is estimated.

The polynomial regression (green line Figure 2) also performed as poorly as the linear regression

which is surprising since a polynomial model fit the theory of the problem better. This lack of fit is likely due to players in long and high paying contracts, not having the incentive to perform very well since they will get paid almost regardless of how they play. Since this model is more complicated and does not perform better than than the linear model, it should be discarded. It's inclusion was only to illustrate that sometimes theory can be plausible, but not necessarily correct.

The fact that models using runs created do 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 and keep the MLB competitive regardless of payroll.

Bibliography

Hall, Stephen, Stefan Szymanski, and Andrew S. Zimbalist. 2002. "Testing Causality Between Team Performance and Payroll." *Journal of Sports Economics* 3 (May): 149–68. https://doi.org/10.1177/152700250200300204.

Hasan, Shahriar, and Thompson Rivers. 2011. "Can Money Buy Success ?: A Study of Team Payroll and Performance in the MLB." www.semanticscholar.org. https://www.semanticscholar.org/paper/Can-Money-Buy-Success-%3A-A-Study-of-Team-Payroll-and-Hasan-Rivers/6f39892055920f0a7855fbc82f50db2f5b46bcf8.

Langs, Sarah. 2022. "Longest Contracts in Baseball History." MLB.com; Major League Baseball.

https://www.mlb.com/news/longest-contracts-in-baseball-history.

Loree, Jacob Andrew. 2016. "Determinants of Baseball Success: An Econometric Approach." *Business and Economic Research* 6 (July): 1. https://doi.org/10.5296/ber.v6i2.9488.

Scully, Gerald W. 1974. "Pay and Performance in Major League Baseball." *The American Economic Review* 64 (December): 915–30. https://www.jstor.org/stable/1815242?seq=4#metadata_info_tab_contents.