Spring 2022

W. A. R. is Worth It

Brandon King
bking19@georgefox.edu

Shane White
swhite19@georgefox.edu

Spring 2022

W.A.R is Worth It

Brandon King

Shane White

W.A.R is Worth It

Brandon King and Shane White

George Fox University

Abstract

This study is evaluating how much a Major League Baseball player should earn in salary based on their WAR (Wins Above Replacement) metric while controlling for a player's veteran status and their current team's market size. All data analyzed in this study was collected from the 2021 baseball season. The population of this study consisted of 213 MLB players that played in at least eighty games throughout the 2021 season. We hypothesized that if a player has veteran status, their team has a large market size, and they produce a high WAR metric that it will result in the player earning a larger salary. Our results did in fact support our hypothesis, and the strongest correlation within our regression analysis was between WAR and the market size of a Major League Baseball team.

Keywords: Wins Above Replacement, Salary, Team Market Size, Veteran Status, 2021 Major League Baseball season, and eighty games.

JEL codes: C55, D61, Z2

Part 1: Introduction

There is an advanced statistic in the game of baseball that can summarize a player's overall worth based on a wide variety of statistics. This metric is called WAR (Wins Above Replacement), and it can encapsulate a player's performance on offense, defense, and how the team does when the player is not in the game. In this study, we looked to answer the question, "How much a Major League Baseball player deserves to make in annual salary depending on their WAR (Wins Above Replacement) metric?". We believe that we can provide baseball fans with a simple explanation of why their favorite player may be earning a salary that is lower or higher then what they actually deserve, and that this can be explained by their below or above-average performance and their age. Also, the variable of team market size factors in the direct effect fans can have on whether a team can afford to keep all-star players because market size has to deal with a team's revenue and how much the city invests into that specific organization. In order to determine if a player deserves the salary they currently are receiving, we expected as WAR increases by one point, the player's salary is expected to increase a certain amount. Here is the equation for calculating overall WAR:

WAR = (Batting Runs + Baserunning Runs +Fielding Runs + Positional Adjustment + League Adjustment +Replacement Runs) / (Runs Per Win)

Each variable within WAR is calculated using different cumulative offensive and defensive statistics. These then are entered into the WAR formula which then combines all of the relative statistics. *Batting Runs* calculates how much the player contributes to his team's runs. If the *Batting Runs* value equals zero then the player is equal with the league average, if the value is above zero then it is better than the average player, and if the value is below zero then they are worse than the league average (Baseball Evolution, 2004). Next, *Base Running Runs* stat is the

amount of runs a player's base running ability adds to the team's total runs (Instructables, 2018). This statistic uses how well the player runs the bases, weighted stolen bags, and weighted ground into double plays. The *Fielding Runs* metric uses the Ultimate Zone Rating (UZR) which "is the defensive metric used in our WAR calculations to measure fielding runs above average relative to the average player at that position" (Fangraphs, 2014). The *Positional Adjustment* variable recognizes that certain defensive positions are more difficult and volatile to the game. For example, a shortstop is a more difficult position and they prevent more runs than a first baseman, and positional adjustment factors the importance of that specific player's position. The *League Adjustment* is a metric used to even out players who play in different leagues, the American League and the National League. Finally, the *Replacement Runs* metric evaluates what would happen if a player was replaced by a bench player or free agent, and this evaluates how a specific player impacts the team's winning percentage. After adding the following variables up, the sum is divided by *Runs Per Win (RPW)* which is the value that converts runs to wins. *Runs Per Win* is the "average number of runs a team needs to score in order to add one win to their total" (Fangraphs, 2012). This allows teams to observe if having certain players in their starting lineup results in a higher amount of runs to be scored which usually results in a win for the team. We excluded pitchers in our regression analysis because the calculation for WAR is different from non-pitchers (position players), and pitchers on average tend to play only every five games. Finally, this means that the positive coefficient of WAR will indicate that salary is positive and constant.

Within our regression equation, we added the following two variables: if the player is of veteran status and a Major League team's market size. We set veteran status as a binary variable where 1 equals veteran status and 0 equals non-veteran status. To qualify as a veteran in the

MLB a player would need to play more than six seasons. Team market size was measured as a categorical variable with 0 representing the bottom 10 teams with the smallest market, 1 being the middle ten teams with a medium-sized market, and 2 being the ten teams with the largest market. We tested a population of 213 MLB players who played eighty games or more in the 2021 season. We did not observe rookies or players who are still on their rookie contract because these players would reflect statistics according to our established variables that would skew data, and they are not yet paid according to their posted statistics. This is due to the issue that in most scenarios these players are being underpaid according to our definition that along with not being rookie standing or still on their rookie contract as determined by the MLB.

Part 2: Data Overview

Our dataset consists of a population of 213 Major League Baseball players that played in eighty or more games with salary being the outcome variable while WAR, veteran status, and team market size are the explanatory variables. The data on MLB players' WAR metric and veteran status comes from *Baseball Reference*. The salary and market size data comes from the database *Spotrac.*

For the outcome variable, salary, we will be running our regression analysis as log(salary) for the purpose of producing a tighter regression. The first explanatory variable, WAR, is being measured with and without a logarithm. The point of producing a regression with log(WAR) allows the growth of salary to remain positive and constant, rather than the regression being positive and increasing. The next variable, veteran status, is a binary variable. The base group, 0, are players that have non-veteran status while the treatment group, 1, are players that are veterans in the MLB. The final independent variable is the player's team market size. The MLB

measures team market sizes as either large, medium, or small. We decided to make team market size a categorical variable. By doing this, we were able to represent team market size on a scale from 0-2 with zero being a small market team, one being a medium market team, and two being a large market team. Using market size allowed us to observe how organizations that generate more revenue and have a larger fanbase can pay players a higher salary.

The summary statistics for our variables were as follows. The average salary for a Major League Baseball player was $7.95 million and the median salary was $5.0 million. The mean WAR metric from our population 2.01 while the median was only 1.8. The maximum salary that was received by a player, Nolan Arenado, in the 2021 season was $35 million, and the highest WAR was 7.3 by Marcus Semien which means Semien made the largest impact on his team compared to any other player in the league.

	Mean	Median	S.D.	Min	Max
TMMarketSize	1.103	1.000	0.8291	0.0000	2.000
WARa	2.011	1.800	2.002	-2.500	7.300
VeteranStatus	0.6995	1.000	0.4595	0.0000	1.000
Salary	7.949e+006	5.000e+006	7.621e+006	1.000e+005	3.500e+007

Although we did observe a population of players, we acknowledge that a population of more than 213 players could help make our data more accurate. Since we conducted research using cross-sectional data, it does highlight some weaknesses that leave our data from truly proving if a player deserves to earn a certain salary. Firstly, since this study is only based on 2021 MLB statistics and salaries, it does not factor in past seasonal performances by players. After all, in professional sports players usually are being paid on how they perform over a course of multiple seasons. This means the data are not completely describing why players are earning the salary they are currently receiving. Also, since we are not observing WAR over a course of multiple seasons, it is hard to say if a player is actually as good or bad as the WAR they earned in

the 2021 season. For example, based on a regression equation, we could determine that a player who earned a WAR of 6.2, who is not a veteran, and plays on a medium market team should receive a salary of $14.89 million. This proves that our regression equation omits the issue like players having a fluke season. Now, this would be an accurate prediction of salary if production and earnings were based on a one-year basis; however, this proves that a stronger study would look at multiple seasons in order to determine what a player deserved based on their overall WAR metric over the course of their career.

If we were to run our regression model without a logarithm on salary (log(salary)) then our regression model would be exponential which can cause our data points to be positive and increasing. By using log(salary), the data became positive and constant. This allows our R^2, coefficients, and standard errors to be more reliable within the data. When we run regressions analysis for log(WAR) against log(salary) our regression is heteroskedastic. Heteroskedasticity is the case in which the errors have a non-constant variance. However, it was beneficial to keep our data heteroskedastic because it produced a larger R^2 value meaning our data was tighter and more closely correlated when we used log(WAR).

Part 3: Methodology

Early on in our project we had trouble figuring out what variables could have significant effects on salary in order for us to evaluate them. Obviously statistics (we referenced WAR as our variable to for statistical analysis) reflect a player's efficiency in terms of winning games but we took a great amount of time looking at some variables that may not seem so obvious that still affect how much a player gets paid. We came to the conclusion of evaluating both veteran status and team market size. For a time we wanted to look at jersey sales but were not able to find the

numbers and also came to the conclusion that this variable only matters to the most popular players in the league that are at the point of making so much money that their jersey sales would be negligible when evaluating their salaries.

Veteran status is our binary variable and this has a large coefficient due to the fact that it has such a large impact on the predicted value of a player's salary. It is not continuous which also impacts the value of the coefficient. Team market size is our categorical variable and has values of 0, 1, or 2. This makes the coefficient that much more important and as we can see leads to a much larger value compared to our coefficient for WAR because WAR is a continuous variable. WAR however is going to have the largest effect because of the much larger range of values that it includes.

A good number of regressions were evaluated and many had positives and negatives that came with each but we ultimately decided to go with a regression that evaluated the logarithm of both salary and WAR while evaluating the variables for veteran status and team market size at base value. This allowed us to interpret the slope for WAR in our regression as positive and increasing. This is a more practical view on WAR and how it correlates to salary because Major League Baseball teams have shown that they are willing to pay significantly more in terms of salary as WAR goes up but the differences in WAR stay the same. WAR is harder to accumulate the higher a given player's WAR value is due to the fact that they must have a very high success rate in all categories ranging from their statistics in all batting categories and defensive categories.

Our prediction equation follows the format below:

$$\log(Salary) = \beta_0 + \beta_1 (log(WAR)) + \beta_2 (VeteranStatus) + \beta_3 (TeamMarketSize) + u$$

Our sample sizes and standard errors for the various coefficients are referenced in part 4.

Part 4: Results & Interpretations

Our original regression was evaluated as level-level given our variables. We started with no logarithms or quadratics as we were looking to gather an understanding of what our data might be telling us. With a one unit increase in W.A.R., ceteris paribus, there is an increase of $980,934 in salary. Upon further observations we found that it might be useful to evaluate our data along with their given quadratic or logarithmic forms.

$$salary = -158{,}170 + 980{,}934(WAR) + 1{,}790{,}010(TeamMarketSize) + 5{,}954{,}650(VeteranStatus)$$

R^2: .268

When evaluating the effects of the logarithm of WAR on salary we did not find much success as this would imply that as WAR goes up then there is a plateau but this is wrong as someone might expect the higher WAR is the more people are willing to pay, and this relationship increases exponentially.

$$\log(salary) = 14.1498 +.202907(WAR) + .280455(TeamMarketSize) + .656185(VeteranStatus)$$

R^2: .227

We then switched to a log-level regression to accommodate for this methodology and it worked out as we would have expected and raised our R^2 by around 5 percentage points. This shows that the regression was aligning well with the quadratic idea that we had proposed. Technically, this is not quadratic but does look to emphasize the importance of WAR as it gets larger and how differences in WAR mean more as they get higher as opposed to lower when it comes to determining salary.

$$salary = -132{,}003 + 2{,}661{,}490(\log(WAR)) + 1{,}743{,}150(TeamMarketSize) + 6{,}715{,}010(VeteranStatus)$$

R^2: .332

The log-log regression we came up with was very helpful as it looked to evaluate the relationship not only of just the logarithms of salary and WAR but also further look into what the differences in WAR mean as they get higher as opposed to lower and how this is much more significant based on player salaries. With a 1% increase in WAR, ceteris paribus, there is a .487582% increase in salary. However, in today's MLB players that are better are being overpaid which is why our regression has a higher R^2 with a logarithmic model.

$$\log(salary) = 14.2788 + .487582(\log(WAR)) + .227752(TeamMarketSize) + .839128(VeteranStatus)$$

R^2: .325

Regardless of what regression was used we found that veteran status had a much greater impact on salary size compared to what we had predicted at the beginning of this research project. This would reflect the decisions made in recent history by Major League Baseball teams to pay their players larger sums of money that many would deem unreasonable, however, it would make sense now because of a presence and the knowledge that they are able to provide to a team that is not measurable in almost all cases.

The limitations to our study are the sample size in that having a larger sample size would help with the accuracy of our regression and the predictions it makes based on the coefficients. In some regressions that we made there was no meaningful interpretation. Not all factors that could have an impact on salary have measurable effects. For example, leadership is something that teams are willing to pay for but this is not measurable. The quality of a teammate is immeasurable as well because there is no number that you can put on being a good or a bad teammate.

From our findings we would expect the quality of Major League Baseball to be much higher. In other words, the effectiveness of the money spent on players would help reflect the teams overall performance. The more a team is willing to spend in terms of money then the

greater success they will find, ultimately boosting sales and possibly the morale of a city which can have a positive impact on the whole population. For example, the Oakland Athletics are arguably the team that is least willing to pay for elite players, and this has led to a dying fan base, to the point that relocating is being heavily considered. On the other hand, the New York Yankees are top spenders every year in terms of player salaries and have found great success with fan loyalty. Not only is player performance a major determinant in salary, but an organization's willingness to pay is also a large factor in how much a player earns.

Dependent Variable: Log of Salary, 213 observations

Regressor	**1**	**2**	**3**	**4**
WAR	.210948*** (.0403874)	.213118*** (.0386414)	.202907*** (.0380825)	
Veteran Status		.759301*** (.167954)	.656185*** (.168428)	.839128*** (.157199)
Team Market Size			.280455*** (.0938551)	.227752** (.0882456)
WAR (logarithm)				.487582*** (.0785062)
Intercept	14.8999*** (.114722)	14.3654*** (.161314)	14.1498*** (.174003)	14.2788*** (.155489)
Summary Statistics				
SER	1.173362	1.122558	1.101851	0.922865
R^2	0.114974	0.193812	0.226996	0.324589

Part 5: Conclusion

Our study evaluated how much MLB players should earn in salary based on their WAR metric from the 2021 season. We decided that this metric would be ideal to use to determine potential salaries of players because it is a combination of a variety of statistics that summarize a players worth to a team. When conducting our regression analysis we found that there was a higher R^2 (tighter cluster of outcomes) when we used log(WAR). Also, team market size and veteran status are factors in how much a player is being payed.

From our research and findings, it is safe to say that having a high WAR metric will earn players a higher salary; therefore, being a well-rounded player pays more in the MLB. This means that players can potentially use their WAR metric to bargain a higher salary using our regression equation because it provides an estimation of a year's worth of salary based on the size of that player's team market, their veteran status, and their WAR metric. The regression analysis provides a straightforward solution to the complex negotiations that occur over a player's salary. Next, this can benefit Major League Baseball organizations because front offices can prove if they are over-paying a player based on the underperformance of an individual's WAR; therefore, providing teams with the capability to use their money more efficiently. Finally, our study illustrates that if a team is willing to spend more money on good players then fan morale and happiness will improve because their team will most likely win more games.

References

Baseball Reference. (2021). Single-season leaders & records for wins above replacement. Baseball. Retrieved April 5, 2022, from https://www.baseball-reference.com/leaders/WAR_season.shtml

Slowinski, P. (2012, April). *War for position players*. Sabermetrics Library. Retrieved April 5, 2022, from https://library.fangraphs.com/war/war-position-players/

Spotrac. (2021). MLB 2022 payroll tracker. Spotrac.com. Retrieved April 5, 2022, from https://www.spotrac.com/mlb/payroll/

Stock, James H., and Mark W. Watson. *Introduction to Econometrics*. Pearson, 2018.