2004

Racial Discrimination in Major League Baseball: Can It Exist When Performance is Crystal Clear?

Will Irwin '04
Illinois Wesleyan University

Follow this and additional works at: https://digitalcommons.iwu.edu/parkplace

Racial Discrimination in Major League Baseball: Can It Exist When Performance is Crystal Clear?

Abstract

American labor markets have a long history of racial discrimination. Human capital theory demonstrates that, on average, workers are paid according to their likely productivity, which is estimated by past experiences in their working lives such as education, previous work experience, and on-the-job training. With MLB, productivity factors that would be included in a human capital theory based equation might include things such as home runs, runs batted in, batting average, years of major league experience, and defensive ability.

Racial Discrimination in Major League Baseball: Can it Exist When Performance is Crystal Clear?

Will Irwin

I. INTRODUCTION

Major League Baseball (MLB) Players make millions of dollars per year to play a game. The players come from all over the world to the United States' largest cities to perform in front of millions of fans each year. From children in the Dominican Republic playing on gravel fields with broomsticks for bats, socks for balls, and milk jugs for gloves to children in the wealthy suburbs of southern California, little boys grow up dreaming of playing for their favorite team. Typically these dreams revolve around hitting a home run to win the World Series or making a diving catch to save a game and, for extremely underprivileged children, the promise of large contracts to bring their entire family out of poverty provide extra inspiration to make it to the big leagues. But are their dreams tainted by the specter of discrimination?

American labor markets have a long history of racial discrimination. Many programs such as affirmative action and equal employment opportunity laws have been instituted to ensure that minority races are not discriminated against during the hiring process, but less has been done to guarantee that once minority workers have been hired that they receive equal pay to their white co-workers. The problem with most industries is that discriminatory firms can

> "When a white player hits 50 home runs it should theoretically be treated the same as when a Latino player hits 50 home runs, which makes for variables that have no implicit bias. "

find reasons to discriminate against minority workers based on productivity factors. Human capital theory demonstrates that, on average, workers are paid according to their likely productivity, which is estimated by past experiences in their working lives such as education, previous work experience, and on-the-job training. Since oftentimes minorities have been discriminated against their entire lives in areas such as these, it becomes very easy for a firm to use these reasons in order to "legally" discriminate against them, which is exactly why racial discrimination is so difficult to measure in normal industries. All of the productivity variables have an implicit racial bias built into them, which blurs the data and makes any finding on discrimination suspect.

With MLB, productivity factors that would be included in a human capital theory based equation might include things such as home runs, runs batted in, batting average, years of major league experience, and defensive ability. When a white player hits 50 home runs it should theoretically be treated the same as when a Latino player hits 50 home runs, which makes for variables that have no implicit bias. Another benefit of using MLB players to assess the state of racial wage discrimination is that many players, including some of the top players in the league, come from the poorest areas in Latin America, which shows that professional baseball is an industry with

little class resistance. In many industries, where advanced levels of education are necessary, it is virtually impossible for a child from a poverty-stricken country to attain employment in that industry. Examples might include doctors, lawyers, accountants, and business executives. Since MLB, as an industry, does not depend on education, but on athletic ability, a trait that can develop in even the poorest of circumstances, it does not fall victim to built in class biases.

By studying racial discrimination in MLB, I hope to show that it does not exist. Since baseball teams that discriminate would only end up hurting their own chances of winning, it seems that there is a strong incentive to provide fair wages. Also, the strong influence of agents on contract negotiations should insure that all contracts are awarded solely on the basis of performance. I hypothesize that racial discrimination does not exist in MLB and that race is not a significant predictor of players' salaries after controlling for productivity.

> "Since baseball teams that discriminate would only end up hurting their own chances of winning, it seems that there is a strong incentive to provide fair wages."

The following section, Section II, discusses some of the past research about discrimination in professional sports. Section III presents my theoretical model, and Section IV develops my empirical model. Section V shows my variables and the regression model. Section VI describes my data sources and offers some critique of their usefulness. Section VII presents my results from the regression equations, and Section VIII discusses the conclusions reached from the results and any possible policy implications that may be needed.

II. LITERATURE REVIEW

The study of discrimination in the work force is rather extensive. Copious amounts of research exist in the areas of racial and gender discrimination. In these studies, racial discrimination is routinely uncovered and shown to exist (Calaway, 1999). Some research even examines MLB. However, none of that discrimination research deals with hitters in MLB.

In an article by Bodvarsson and Pettman (2002) entitled "Racial Wage Discrimination in Major League Baseball: Do Free Agency and League Size Matter," pitchers are used to determine whether or not racial wage discrimination exists. This study uses a logarithmic regression model with variables such as race (white or non-white), population of the player's metropolitan area, productivity variables, and racial make-up of the player's metropolitan area. Since the study dealt with pitchers, the performance variables are very different. Examples include: wins, losses, saves, strikeouts, and earned run average. The study attempts to determine whether or not the addition of more teams to MLB would eliminate any racial discrimination that existed. The theory was that more teams would increase the level of competition between teams for good players, thereby causing teams to offer fair wages to get the best players. The results of this study show that expansion in 1993 did indeed eliminate discrimination for pitchers.

In a study titled "Equity and Arbitration in Major League Baseball," Fizel, Krautman, and Hadley (2002) look at a theory called "Equity Theory," which states that workers compare their efforts to their co-workers that are in comparable situations and then adjust their behavior based on that comparison. They hypothesize that MLB players look to players of similar abilities to decide how much they should earn when they file for arbitration and free agency. They use both regression and bivariate techniques to determine whether or not a player will decide to file for arbitration. The results indicate that players do look at their co-workers performance and salary when they apply for arbitration. They do not include a measure of race in their study. If they had, perhaps they might have found whether minority players make different decisions than white players on this topic.

In the context of the current literature, it is

extremely important to test for the presence of discrimination amongst MLB hitters. Since no such research is currently available, it is imperative to analyze this issue to ensure that minority players are being treated fairly.

III. THEORETICAL MODEL

The theoretical basis for this research is human capital theory. This theory aims to explain how productivity factors of a worker's history influence how much they are paid in their jobs. Human capital is seen as a form of investment. People can invest in schooling in order to increase their level of human capital. Similarly, firms can invest in on-the-job training for their workers. Since education increases a worker's knowledge for skills that can be applied to a job, theory suggests that this increase in knowledge will allow the worker to be more productive, which will give them the ability to bargain for higher wages (Rima, 1981). Firms will pay higher wages since the marginal productivity of their business will be increased with the addition of more human capital to their workforce. Also, since investing in human capital can have costs, such as tuition for schooling and the opportunity cost of lost wages during the time period of schooling, employers can be forced to pay higher wages to people with more human capital (Mincer, 1958).

In a traditional work setting, the factors that go into human capital would include educational attainment level, years of work experience, on-the-job training, and age. All of these variables, theoretically, should give an indication of how productive a worker will be. Once an employer knows how productive the worker will be, he/she can pay them a wage that is fair for that level of productivity.

For baseball players, the factors that employers look at to determine salary are very different from the traditional job market. Athletic talent is the main form of human capital for baseball players. Since players are simply born with this talent, they are not rewarded for any investment they make into their human capital, but rather they are compensated for their natural ability to produce in an athletic industry. However, just as job experience is a form

of human capital for traditional industries, the amount of time playing high-level professional baseball should be an important human capital factor for baseball players.

Therefore, a human capital model for baseball will have some different factors, but the factors in the model will represent productivity in the same way that education, experience, and training do in a traditional human capital model.

IV. EMPIRICAL MODEL

The equation for a traditional human capital model follows:

Wage = f (education, work experience, on-the-job training, age)

Notice, a factor not included amongst those variables is race. When studying race as a determinant of wage compensation, it is important to account for individual differences in productivity due to the factors mentioned above in order to determine whether or not race has an effect. Since, theoretically, race alone should not affect a person's productivity, it would be expected that different races would receive equal pay, *ceteris paribus*. Unfortunately, this is not always the case.

In order to determine whether or not MLB players are receiving fair pay, I include race as one of the determinant variables. Other determinant variables account for traditional productivity factors. These include measures for offensive and defensive ability, experience, and reliability of future production.

It is important to note that MLB teams are located in major cities across the country and vary widely in size from New York City, with a metropolitan area of over 20 million people, to Milwaukee, with a metropolitan area of just over one million people (U.S. Census, 2003). Since teams in smaller cities would have to pay just as much as teams in large cities to attract players, some may say that it is doubtful that a team in a small market can pay a player equally well as a large market team due to the large differences in total revenues. However, for the purpose of this study, it is assumed that a player would

not accept a lower salary to play in a small market, and, therefore, that all salary figures are true values awarded for the level of productivity that is expected of the player. Given this assumption, I include no measure of market size in the model.

Conceptually, a players contract value could be thought of depending on several categories of variables:

*Contract Value = f (offensive ability, defensive ability,
experience, reliability, race)*

From this theoretical equation, I build a regression equation that I use to test my hypothesis that racial discrimination does not exist in MLB.

V. VARIABLES AND REGRESSION

The regression model incorporates the variables described in the theoretical model and applies them specifically to the research problem by using the available data.

The dependent variable is salary. The data for this variable consists of a one-year contract value, the first year of a new contract. The number of years of the contract is not included. This is unfortunate because length of contract could be another variable in which discrimination could play a role. However, data for length of contract is not readily available.

The regression equation attempts to account for different aspects of a player's abilities. For the purpose of limiting the data and eliminating positional biases, I use only outfielders in this study. Typically, scouts and general managers judge players on the basis of "five tools": hitting for average, hitting for power, ability to steal, ability to catch, and ability to throw. Table 1 presents the variables.

A. Offensive Statistics and Variables

Typically, traditional statistics like batting average, runs batted in, runs scored, home runs, and steals measure offensive production. The problem with using such statistics is that they do not fully capture the ability of all players to contribute to their teams. For example, Player A plays for a team in which the players in front of him and behind him in the line-up all hit for high average and power, while Player B plays for a team in which the players immediately surrounding him in the lineup all hit for poor average and power. If Player A and Player B have exactly the same inherent offensive capability, but are playing with these different teammates, Player A will have more runs scored and more runs batted in because there will be more men on base ahead of him to drive in and better hitters behind him in order to drive him in. Some suggest that this effect will only bias the runs scored and runs batted in measures of performance, but it is generally assumed and widely accepted that home runs and batting average can also be positively and negatively influenced depending on the players surrounding a player in the line-up. In order to insure that statistical studies of baseball hitters are not flawed due to this "team influence," a statistic has been created by baseball scholar Bill James called "Runs Created" (Dunning, 2003).

The formula for Runs Created has numerous variations. It has been changed many times by its creator and also by other researchers that have used it. For this study, the most basic formula is used due to the availability of data (Baseball-Reference, 2003).

I use the following formula:

*Runs Created = (Hits + Walks) * Total Bases / Plate
Appearances*

"Total bases" represents the number of bases reached by a player due to hits. Thus, for a single, total bases equals one, for a double, total bases equals two, for a triple, total bases equals three, and for a home run, total bases equals four. "Plate appearances" represents the total number of times a player comes to the plate. It differs from "At bats" because "At bats" subtracts out walks, sacrifices, and hits-by-pitch. "At bats" is the statistic used to calculate batting average (Sports, 2003).

Runs Created attempts to estimate a player's total contribution to his team's total output. It aims to eliminate the "double-counting effect" that occurs when a team scores a run. To explain, each time a

run is scored, the player who actually crosses the plate gets credit for a "run scored", while the player that got the hit to drive the other player in gets credit for a "run batted in." Statistically, the run is counted twice. Therefore, the players on a team that scores 700 runs in a season might have a sum of 1400 runs scored and runs batted in (minus special cases where runs score without a run batted in being awarded). In order to measure which players on each team are actually most responsible for the total offensive output, the Runs Created variable gives one player more credit than the other when each run is scored (Dunning, 2003). The following hypothetical situation further illustrates this point:

> *Player A hits a triple to lead off an inning. Player B sacrifices Player A home with a long fly ball to centerfield. In this situation, the statistical breakdown would be: Player A – 1 hit in 1 at bat for a 1.000% batting average and 1 run scored, Player B – 0 hits in 0 at bats for no change to his batting average and 1 run batted in (sacrifice fly outs do not count as an official at bat). If these players' statistics are compared, Player A would only receive a slight advantage over Player B in the batting average category. Their run production would be considered equal.*

This example of traditional stat-keeping shows how the player that "did the most" in scoring the run does not earn the credit that he might technically deserve. Runs Created solves this problem by giving "extra credit." Player A would receive three total bases and Player B would receive zero total bases.

For regression analysis, Runs Created appears to be a good alternative to traditional statistics because it eliminates some of the team bias that is implicit in the data. Also, by reducing the number of independent variables in the regression, using Runs

Created saves valuable degrees of freedom.

In order to further normalize the variable, each players' total Runs Created, for the years before they sign their contract, is divided by the number of games they have played and then multiplied by 162 (the number of games in a full season). This should show their average ability to produce offensively.

The ability to steal is measured with a stolen base statistic. The same procedure used with Runs Created is also used with steals. Each player's steals value shows the number of steals they have had on average per season. On a side note, this statistic could be extremely interesting due to the fact that some players' main purpose for being in a line-up is to steal bases. It will be interesting to see if these stealing specialists are rewarded for their rare talent.

B. Defensive Variables

Defensive tools, such as the ability to catch and to throw, are more difficult to quantify. For the ability to catch, the most commonly used statistic is fielding percentage, but the problem with this statistic is that some slower players do not have the opportunity to attempt to make difficult plays and therefore end up having higher fielding percentages than the faster, flashier fielders. This problem is virtually unavoidable with the data, so I do not include fielding percentage in the regression.

The ability to throw can be easily quantified with outfield assists. An assist is awarded to a player when they throw the ball to another player and this action results in an out, force or tag. For outfielders, this statistic is extremely important to their overall defensive ability. For the model, assists are included as a 162-game average, calculated in the same way as the offensive statistics described above.

C. Experience Variables

In order to account for the traditional human capital theory variable of experience, I include the number of games played in MLB. Since many players play a few games in a year before playing their first full year, it would be impossible to include

a variable such as years of experience. Number of games played will clearly differentiate between young players and veterans, while also allowing all available statistics prior to contract signing to be used in calculating potential productivity.

An element of productivity that has nothing to do with the game of baseball, but will still have an impact on the human-capital thought processes that general managers have when deciding how much money to give to a player is age. Oftentimes, a team awards a younger player a large amount of money over a long period of time, while an older player may receive a one or two year contract worth an

These variables may give some indication of reliability, as was discussed before. One can assume that an older, more experienced player will have had a longer period of time to demonstrate exactly how good of a player they are. A general manager may be willing to pay more if they have a deeper, more concrete understanding of what they are getting for the money they are paying. On the other hand, a young player could be considered more valuable due to his youth and potential to develop. However, it seems that baseball salaries are awarded more on past performance than on potential future performance, giv-

TABLE 1
Variables and Descriptions

Variable	Description	Mean	Maximum	Minimum	Standard Deviation
Salary	Value of Contract	2,440,064.1	7,550,000	250,000	1,985,099.9
RC (+)	(Runs Created/GP)*162	78.1	125.6	39.9	20.9
ASSIST (+)	(Assist/GP)*162	9.1	16.7	2.8	3.8
GP (+)	Games played before contract was awarded	523.5	982	168	208.7
AGE (+ or -)	Age on July 1st of the year contrcat was awarded	27.9	34	23	2.3
BLACK (0)	1 = Black, 0 = Other	.37	1	0	.486
LATINO (0)	1 = Latino, 0 = Other	.33	1	0	.474

extremely large amount of money. Since no research exists on how age affects salary awards in MLB, it is uncertain what effect age will have in the regression equation. However, in order to receive a more accurate result from the regression, age must be accounted for with an independent variable.

ing the advantage to the older, more established players.

D. Racial Variables
I use dummy variables to capture the effect of race on player salaries. For both African-Ameri-

can and Latino players, a one or a zero is used to designate a player in either of those racial categories. I capture the effect of being Caucasian using an omitted variable. Placing players into the aforementioned categories was accomplished using country of origin and other factors.

Resulting from the variables in Table 1 will be a regression equation:

Player Salary = f (RC, ASSIST, STEALS, GP, AGE, BLACK, LATINO)

VI. DATA

The wonderful thing about working with MLB data is that people are fanatical about keeping track of statistics about every aspect of the game. From offensive statistics to defensive statistics, to player salaries and free agent filings, all of the data found in this paper is readily accessible on reputable and reliable websites. It is important to note the ease of Internet research when working with baseball statistics. Efficiency is drastically increased due to the extreme quickness by which player data can be found on these websites. Additionally, definitions of statistics and explanations of how to interpret them are found easily on the Internet.

The data for offensive and defensive statistics is available from *Baseball-Reference*. This website is basically an encyclopedia of baseball players and has records of almost every player that has ever played in the major leagues.

Data for player salaries can be found at a website run by the chairman of the Business of Baseball Committee of the Society for American Baseball Research. The site is particularly useful in determining when the players received their salary payments as a result of a free agent signing because it has a list of all free agent filings made by players. Using this list, the year in which the player entered the free agent market can be found and their performance data from the previous years can be used to determine on what basis their new contract was awarded.

The data for the race variable is less concrete. In many cases the player's country of origin indicates which category they should be placed in (Baseball-Reference, 2003). For example, all play-

ers from the Dominican Republic are considered to be Latino. The same can be said for players from any other country in Central and South America. For minority players born in the United States, the distinction is more difficult to make. Typically, a determination can be made based on the origin of the player's last name. A name of Hispanic origin would tend to indicate that the player is Latino. Conversely, a recognizable "American name" would categorize the player as black. Any cases where no determination could be made with some certainty are excluded.

VII. RESULTS

The results from the regression are quite good. The model has an Adjusted R-square value of .545, showing a fairly high goodness of fit.

As for independent variables, the unstandardized coefficients, t-statistics, and significance values are presented in Table 2.

The two significant variables are Runs Created and Games Played. The rest are not significant with extremely low significance values.

Runs Created has a significance value of .003 and, as expected, a positive coefficient of 43,128, indicating that an additional run created results in 43,128 more dollars in the value of a player's contract. To put this in perspective, adding a Runs Created value of 10 for a season would add $430,127 to a player's salary. The average Runs Created value for the sample is 78.2. Traditionally, a season of 100 runs created is considered a very good season.

The other significant variable, Games Played, has a positive coefficient of 5,448 and a significance value of .000. This variable indicates that an additional year of experience (162 more games played) should translate, on average, into 882,511 more dollars. Obviously, an upper limit to this variable must exist or the oldest players in the league would all be the highest paid, but for this study, the positive value of the coefficient was expected. Since the group of players used in this group is relatively young, their games played numbers are also relatively low, so it is to be expected that more ex-

perience will benefit these younger workers. The highest number of games played in this study is 982, so these results can only indicate a rate of return on games played up to that point. After that, or possibly before, the return to games played may become negative, reflecting the older age of the player, which may make a player less desirable.

Steals, Assists, and Age are all non-significant variables in this regression with significance values of .977, .753, and .411, respectively. All of these variables were expected to have positive coefficients, but their low significance values suggest

that the signs on the coefficients are relatively meaningless. According to these results, steals, defense, and age are not important determinants of baseball players' salaries.

Most important for the purpose of this study are the race dummy variables, both of which are not significant. The dummy variable for blacks carries a positive coefficient of 555,729.0 with a significance value of .296, while the dummy variable for Latinos carries a positive coefficient of 202,386.1 and a significance value of .691. Since the variables are not significant, the signs and magnitudes of the coefficients are not important.

VIII. CONCLUSION

The purpose of this study is to determine whether or not racial discrimination exists in MLB. After running the regression equation and finding race variables that were not significant, it can be concluded that under this set of circumstances, racial discrimination does not exist for outfielders in MLB. The positive implications of this are obvious. It is an indication that if productivity is clear and unbiased in nature, employers will provide fair wages to all employees. Obviously, very few industries exist with such clear cut and easily measured productivity, but just the fact that minority workers are paid fairly when productivity is clear is a positive sign for the state of the American social consciousness.

Future research could examine total contract value, which would capture the length of contracts to see if baseball owners prefer one race to another for long term contracts. Additionally, researchers could expand the regression to include all positional players, to try to get a better representation of the entire league's salary structure.

TABLE 2
Regression Results for Salary

Variables	Coefficient	Standard Error
Constant	-1,319,631.0 (-.406)	3,248,449.0
RC	43,127.6* (3.171)	13,601.4
ASSIST	-21,341.3 (-.316)	67,442.0
STEALS	-520.7 (-.029)	17,834.0
GP	5,447.6** (4.114)	1,324.3
AGE	-90,453.0 (-.830)	108,941.0
BLACK	555,729.0 (1.058)	525,330.0
LATINO	202,386.1 (.400)	506,230.1
Adjusted R² = .545 n = 52		

* indicates significance to .01 level
** indicates significance to .001 level

NOTE: t-statistic appears in parentheses

REFERENCES

Baseball-Reference. 2003. 20 November 2003. Available <http://www.baseball-reference.com>.

BaseballStuff. 2003. 20 November 2003. Available <http://www.baseballstuff.com/ fraser/

gloss.html>.

Bodvarsson, Orn B. and Shawn P. Pettman. "Racial Wage Discrimination in Major League Baseball: Do Free Agency and League Size Matter?" <u>Applied Economics Letters</u>, 9, 2002: 791 – 796.

Calaway, Jaynanne. "The Gender Pay Differential: Choice, Tradition, or Overt Discrimination?" Senior Research Honors, Illinois Wesleyan University, 1999.

Dunning, Rodney. 2003. 20 November 2003. Available <htpp://www.wfu.edu/users/dunninrb/ stuff/baseball/runs_created.html>.

Fizel, John, Anthony C. Krautmann, and Lawrence Hadley. "Equity and Arbitration in Major League Baseball." <u>Managerial and Decision Economics</u>, 23, 2002: 427 – 435.

Mincer, Jacob. "Investment in Human Capital and Personal Income Distribution." <u>Journal of Political Economy</u>, Vol. 66, 4, 1958: 281 - 302.

Rima, Ingrid H. "Labor Markets, Wages, and Employment". New York: W.W. Norton & Company, 1981.

Sports Wired. 2003. 20 November 2003. Available <http://www.sports-wired.com/ content/ glossary.asp>.

U.S. Census Bureau. 2003. <u>U.S. Census Bureau</u>. 20 November 2003. Available <http:// eire.census.gov/popest/archives/metro/ma99-02.txt>.