Does Batting Performance Predict Baseball Player Salaries?

Daniel Beauchaine[1]

[1] Portland State University, Portland, Oregon

E-mail: beauc@pdx.edu

Abstract

Baseball fans and teams alike have developed a much broader understanding of player value. I therorize that modern baseball teams maximize their value in terms of this new understanding by signing better players to lucrative contracts. I found that wRAA, a proxy for player value, is statistically significant in the prediction of real player salary. I found that the coefficients associated with wRAA have increased over time. Additionally, I found evidence that this effect was stronger in players at or over the age of 30 than players younger than 30. Specifically, I found that wRAA is not statistically significant in the prediction of real salaries of players under 30.

Keywords: age, baseball, performance, salary, wRAA

1. Introduction

At the top end, Major League Baseball contracts are becoming longer and more lucrative. On the bottom end, young players on rookie contracts are limited by team control and arbitration. Teams are spending more money per player. Over the period 2000-2016, both average and median player salaries have steadily risen in both nominal (figure 1) and real (figure 2) terms. The most lucrative contracts are earned in free agency, where teams compete for the services of players elite or otherwise. Teams, increasingly, spend more time and money developing metrics to differentiate player value and maximize their value per dollar spent.

Batting statistics have evolved with the sport. Originally players were compared to one another through counting stats found in box scores. These simple measures include runs scored (R), runs-batted-in (RBI), and batting average (BA). These are easily observed and calculated and make sense as a starting point. However, these counting stats do not isolate a player's production from that of his teammates. For instance, for a player to score a run, a teammate must drive him in, for a player to drive in a run, there must be a teammate on base ahead of them.

It is true that total runs scored is a good indication of a team's offensive performance. With this in mind, if a team is to maximize their production on the field, they should attempt to maximize individual run production per player. There have been many inquiries into the optimal metric for discerning individual run production. If teams are aware of these new metrics and they are spending efficiently, a team's spending per player should be highly correlated with these metrics.

2. Literature Review

In their article *An Evaluation of Major League Baseball Offensive Performance Models,* Jay Bennet and John Flueck (1983) set out to compare various offensive performance models. This is a somewhat dated article, considering most of our modern understanding of baseball value in the form of sabermetrics has been developed since this paper

was written. Nonetheless, this paper is valuable. The model set forth by Bennet and Flueck (1983) compared various standard offensive statistic models against actual run production. They used a process of systematic linear model-building to determine the best-fit model. The model that Bennet and Flueck (1983) determined best predicted actual run generation is the Expected Run Production model as follows.

$$ERP = 0.499 * H1B + 0.728 * H2B + 1.265 * H3B + 1.449 * HR + 0.353 * TBB + 0.362 * HP + 0.126 * \text{SB} + 0.394 * \text{SF} - 0.395 * \text{GIDP} - 0.085 * \text{OUT} - 67.0$$ (Bennet and Flueck, 1983, p.78)

Bennet and Flueck, by systematically adding statistics one by one, found that the offensive statistics contained in ERP explain 94.76% of the variation in runs scored (r-squared.) Interestingly, they show that adding statistics such as SH (sacrifice hits) and CS (caught stealing) reduced the R-Squared value and determined they should be left out. (Bennet and Flueck, 1983, p.78-79) The ERP model is a precursor to the modern wOBA, which weights each value differently each year.

In their paper *Do Baseball Players Regress Toward the Mean,* Teddy Schall and Gary Smith conclude that "players who do exceptionally well in any particular season typically do not do as well the subsequent season." (Schall and Smith, 2000, p.231) In other words, they regress toward the mean. Schall and Smith (2000) used both pitcher and hitter data. They created a standardized measure of performance Z, which represented the number of standard deviations above or below the mean a particular player was in a given year. Schall and Smith found a statistically significant correlation from one season to the next for both batting average (.37) and ERA (.22). (Schall and Smith, 2000, p. 234) This suggested that year to year performance is not completely random, but also not as predictive of future success as they could be. They compared the standard Z values to rZ values that they adjusted by calculating the adjacent season performance correlation coefficient for all players and multiplying it by the Z value. Using Root mean squared error, they determined that this deflated value rZ was a better projection of the next season's performance than the Z value alone. (Schall and Smith, 200, p.234-235) This paper raises an important point in player valuation; contracts are set based on past performance and expected future performance, Schall and Smith have shown that this is a less than trivial task.

In *Does the Baseball Labor Market Properly Value Pitchers?* John Charles Bradbury (2007) shows that common baseball statistics, such as earned run average (ERA), do not properly isolate the value of pitchers. Specifically, Bradbury shows that batting average on balls in play (BABIP) and fielder quality are included in ERA and muddy the overall quality of value isolation. Bradbury used statistics from the same Lahman database that I am using, including a "pitcher park factor," a deflator used to control for the differences in home park environments. In his regression on ERA Bradbury also controlled for pitcher age, a proxy for experience, used a dummy variable for league, to control for effects of the designated hitter, and controlled for seasonal competition by including a variable for year. (Bradbury, 2007, p. 619) His overall conclusion is that ERA is a bad estimator of pitcher value, since it is highly polluted by BABIP. He estimated that a one standard deviation change in BABIP accounted for a one-half standard deviation change in ERA. (Bradbury, 2007, p.619) Bradbury concludes that it is difficult to isolate pitcher performance from jointly produced outcomes with fielders and batters. This analysis highlights the need for modern metrics which control for factors outside of the player's control.

In their article *Predicting Run Production and Run Prevention in Baseball: The Impact of Sabermetrics,* Phillip Beneventano, Paul D. Berger, and Bruce D. Weinberg (2012) use stepwise multiple regression analysis to assess the run-scored predictive value of various sabermetric (advanced baseball statistic) and traditional statistic models. The stepwise regression chose the sabermetric weighted on base average (wOBA) as the most significant predictor. wOBA alone accounts for nearly 90% of the variation in runs scored (r-squared = .896.) wOBA was followed by % of plate appearances ending in a strikeout (K%), another sabermetric, and two traditional statistics slugging

percentage and on base percentage. They combined these four statistics into the following regression model which produced an r-squared value of .953.

$$\hat{Y} = -903 + 2226 * wOBA - 184 * K\% + 1116 * SLG\% + 1501 * OBP$$

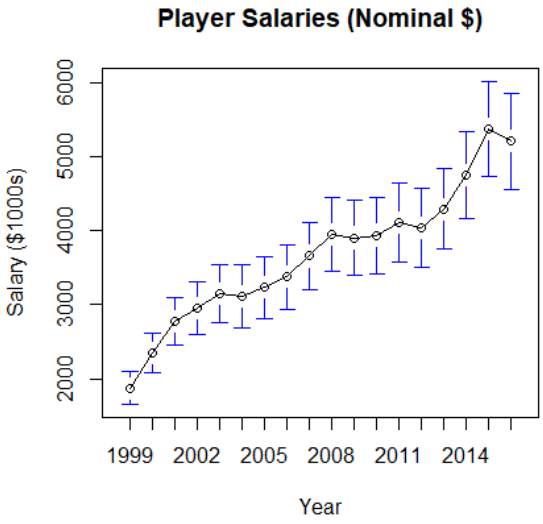
This model represents a combination of sabermetric statistics chosen in order of predictive value. wOBA and K% provided an R-squared value of .930 on their own, with the remaining .023 provided by the traditional statistics. This is evidence that modern sabermetric methods are more valuable predictors of runs scored than traditional statistical methods. (Beneventano, Berger and Weinberg, 2012, p.72) This paper supports my decision to use wOBA as the primary predictor of individual player contribution.

Heather M. O'Neill addresses the tendency of different prediction models to give different, or even opposite, results in *Do Major League Baseball Hitters Engage in Opportunistic Behavior?* (O'Neill, 2013) O'Neill was interested in studying whether Major League Baseball hitters overperform expectations in a contract year (the year before free agency.) She acknowledges that past research has shown that OLS estimations yield contrary results to fixed effects models. (O'Neill, 2013, p.1) O'Neill runs a population regression for players with OPS+ as the dependent variable–her desired measure of performance–with age as a quadratic, games, position as a dummy variable, playoff as a dummy variable, retire as a dummy variable, and contract year as a dummy variable. In the results from this regression show a negative correlation between contract year and OPS+, seemingly contradicting the widely-held belief that players perform better in a contract year (O'Neill, 2013, 13.) She then runs a one-way fixed effects model, with a dummy variable for each player on the same data, controlling for time-invariant traits affecting OPS. The result of the FE model were nearly double R-Squared values compared to the OLS estimation. The FE model shows a 3.98% increase in OPS+ relative to the mean for players playing on a contract year, the opposite conclusion to the OLS estimate. (O'Neill, 2013, 15) In conclusion, O'Neill accounts for the difference in models as the effects of omitted variable bias in the OLS model. She concludes that the FE model controls for unobservable traits that are expected to be correlated with OPS+, therefore yielding more accurate results. O'Neill chose OPS+ as the metric of focus in her study. OPS+ is similar to wOBA, they are both weighted based on external factors. I prefer wOBA since it gives even more granular weight to each of the components of OPS.

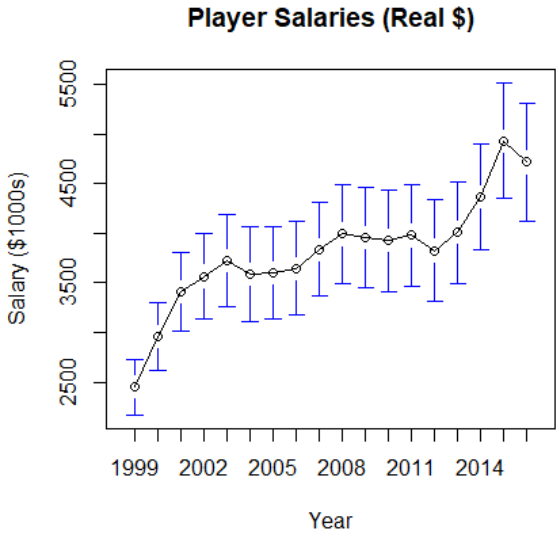
Jay Tymkovich studied whether position on prospect rankings lists was predictive of future success in his thesis *A Study of Minor League Baseball Prospects and Their Expected Future Value.* (Tymkovich, 2012) Tymkovich chose wins above replacement (WAR) as his preferred measure of player success and as his dependent variable. He created separate regressions for three different prospect lists, Baseball America, Baseball Prospectus, and John Sickels. He used WAR as his dependent variable and the number ranking from each list as well as dummy variables for position, league (MLB, AAA, AA, etc..), Team, and Age. Tymkovich found that for every 10 spots higher on Baseball America's top 100 list a player is ranked, he is expected to produce .7 more WAR in his first five major league seasons. He also found that playing in college is associated with 1.5 more WAR over the same period. (Tymkovich, 2012, 31) Overall, his conclusion was that baseball prospect rankings are somewhat predicted of real future value.

3. Trends in Salary

It is unsurprising that player salaries tend to rise over time. Wages in general rise as a result of inflation. I was interested in observing the impact of batting performance on both real and nominal salaries. Player salaries have risen in both nominal and real terms (figures 1 and 2.)



(Figure 1. Nominal Average Player Salary Per Year)

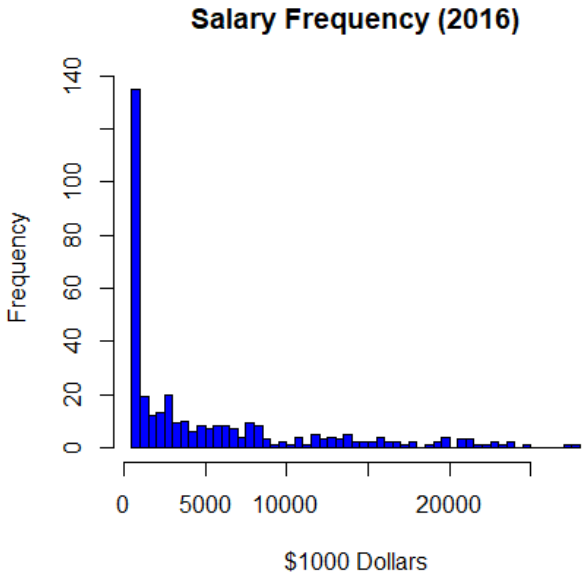


(Figure 2. Real Average Player Salary Per Year)

Major League Baseball salaries are highly skewed due to the fact that all young players are subject to salary limiting stipulations. The 2017 Major League Baseball collective bargaining agreement states

"Following the completion of the term of his Uniform Player's Contract, any Player with 6 or more years of Major League service who has not executed a contract for the next succeeding season shall become a free agent"

A player is under team control until this requirement has been met. A team may pay that player the league minimum over this period unless another salary is negotiated between the team and the player in arbitration or otherwise. Most players achieve their peak performance in their mid to late twenties (Schulz, Musa, Staszewski, & Siegler, 1994.) Many players do not reach free agency until they have reached or passed their prime, and many others never reach it due to injury. The result of the free agency policy is that the majority of contracts are on the low end with a few very valuable contracts on the high end. (figure 3.)



(Figure 3. 2016 Salary Frequencies)

4. Shaping the Data

4.1 Transforming the Lahman Database.

Sean Lahman[1] provides an up-to-date database of the basic box score statistics in baseball for all of recorded baseball history. His database was the starting point for my analysis. This database, however, contains only the base stats (table 1.) Any additional statistic, from average to wRAA have to be calculated from the base statistics. Lahman keeps detailed records over many spreadsheets, my first task was to compile all of the relevant statistics into

[1]"Download Lahman's Baseball Database – SeanLahman.com,"

one dataset. I combined segments of his batting spreadsheet with the salaries and people spreadsheets to create one database with all of the information I needed to calculate player age, performance, and salary. I am interested in the modern era, and therefore have limited my dataset to the years 2000-2016.

TABLE 1— LIST AND DESCRIPTION OF PROVIDED LAHMAN STATS,

Description	Abbreviation
Games Played	G
At Bats	AB
Runs Scored	R
Hits	H
Doubles	2B
Triples	3B
Home Runs	HR
Runs Batted In	RBI
Stolen Bases	SB
Caught Stealing	CS
Walks	BB
Strikeouts	K
Intentional Base on Balls	IBB
Hit by Pitch	HBP
Sacrifice Hits	SH
Sacrifice Fly	SF
Ground into Double Play	GIDP

Lahman's database breaks player seasons not only into years, but stints with individual teams. Many players had multiple entries per year. I was interested in looking at only one observation per player per year. In order to generate the entries I was looking for, I looped through the observations and added all stints per player per year together, removed the duplicate entries and any entry without salary information. I had 14,582 individual observations, each representing a player's performance in a year. To complete the traditional statistical makeup of my dataset I calculated the missing values: singles, average, on-base percentage, slugging percentage, on-base plus slugging, and batting average on balls in play using the formulas from (table 2.)

TABLE 2— FORMULAS FOR TRADITIONAL STATISTICS

Description	Statistic
$H - 2B - 3B - HR$	Singles
H/AB	AVG
$(H + BB + HBP)/(AB + BB + HBP + SF)$	OBP
$(1B + 2 * 2B + 3 * 3B * 4 * HR)/AB$	SLG
$OBP + SLG$	OPS
$(H - HR)/(AB - K - HR + SF)$	BABIP

4.2 Weighted Runs Above Average.

Weighted Runs Above Average (wRAA) is a top predictor of runs produced and runs produced are a top predictor of team wins. Other statistics, such as WAR, attempt to measure team win contributions by individuals directly. WAR, however, requires access to subjective statistics such as ultimate zone rating (UZR) which require video analysis to produce. WAR cannot be derived from Lahman's database. wRAA is derived from weighted on-base percentage (wOBA.) wOBA is a weighted average of the traditional components of on-base percentage (walks, hit by pitch, singles, doubles, triples, and home-runs.) Each of these outcomes is assigned a weight based on their relative value. These weights signify the fact that not all hits (or walks) are created equal; certainly, a triple should be worth more than a single.

A simpler technique for answering this question is on-base plus slugging (OPS.) OPS is a sum of the slugging percentage, which weights the various hit outcomes, and on-base percentage, the simple percentage of times a batter reaches base. The problem with this approach is that it assumes that slugging percentage and on-base percentage should be assigned equal weight. It also assumes that each outcome should be assigned the same weight every year. wOBA does not, wOBA assigns year-specific weights to every outcome. wRAA is an extension of wOBA that assigns a number of runs added (or subtracted) from a player's individual offensive contributions. wRAA further weights a player's contributions against league average wOBA. This makes wRAA slightly preferable to wOBA as a metric.

Calculating wOBA requires access to data that is not available in the Lahman database. I downloaded a spreadsheet of wOBA information from Baseball Reference[2]. This file contains the wOBA weights required to calculate wOBA and wRAA from Lahman's database. These weights are denoted by a w in the equations in (table 3.) wOBA weights differ per season, I matched each observation to the wOBA values for the observation year to create a wOBA and wRAA value for each observation.

[2]"wOBA | FanGraphs Baseball,"

TABLE 3— FORMULAS FOR ADVANCED STATISTICS

Description	Statistic
$\frac{wuBB * uBB + wHBP * HBP + w1B * 1B + w2B * 2B + w3B * 3B + wHR * HR)}{AB + BB - IBB + SF + HBP}$	wOBA
$\frac{wOBA - league\ wOBA}{wOBA\ Scale} * PA$	wRAA

4.3 Further Transformation

Major League contracts are not always re-evaluated annually. To control for the effect of multi-year contracts on salary, I created a lag variable containing the previous year's salary for each observation. For example, if the observation is in year 2010 the lag variable would contain the individual's salary value from 2009. Additionally, I restricted my set of data to individuals with 100 or more at bats to eliminate observations with very small sample sizes and pitchers.

To control for the fact that wages, including Major League salaries, rise with inflation. I downloaded the nominal yearly CPI from the World Bank[3] online database and adjusted each observation based on the corresponding yearly nominal CPI. I also created a lag variable for last years real salary. The difference between the relationship between real salary and nominal salary and time can be observed by comparing (figures 1 and 2). In addition, I generated a quadratic term for age.

I isolated data from 2000 to 2016, I was interested in testing whether batting performance has more or less of an impact on salaries over time. I split my observations into separate datasets based on year. For example, the dataset for 2000 contains all of the observations from the full dataset that occurred during the 2000 season.

5. Method

5.1 Performance and Salary by Year

The first relationship I was interested in testing was the relationship between performance, as demonstrated by wRAA, and real salary. In order to test this relationship, I developed the following OLS model.

$$\widehat{realSalary} = \beta_0 + \beta_1 * wRAA + \beta_2 * age + \beta_3 * age^2 + B_4 * lastSalary + u$$

With the quadratic age variable to control for the effect of age on real salary and the previous years salary to control for multi-year contracts. I ran this regression on the years 2000, 2008, and 2016 individually. The results of these regressions can be viewed in (table 4.) In 2000 I had a database of 340 observations with an adjusted R-Squared value of 0.863. In 2008 I had a database of 290 observations with an adjusted R-Squared value of 0.864. In 2016 I had a database of 269 observations with an adjusted R-Squared value of 0.865. These values seem to indicate that this model predicts roughly the same amount of the variation in real salary over these three years, approximately 86%.

A Breusch Pagan test for heteroskedasticity resulted in a Chi-square critical value of 46.22 and a p-value of $1.06 * 10^{-11}$. This strongly suggests that the data is heteroskedastic, I, therefore, included robust standard errors in the regression results. The coefficient for wRAA is 14.439 for 2000, 34.928 for 2008, and 39.590 for 2016. These results are very statistically significant at the 0.001

[3] "World Bank Group - International Development, Poverty, & Sustainability,"

level. This result suggests that player performance, as demonstrated by wRAA, does affect real player salary. wRAA can be either negative or positive and ranges from approximately -40 to 90, suggesting that wRAA can positively or negatively affect player Salary.

Over the period 2000-2016 the coefficient for wRAA on real salary has increased. This suggests that more modern contracts are more aligned with actual batting value than older contracts. This trend could be partially explained by a revolution in sabermetric evaluation of players. Teams, increasingly, have more advanced metrics for evaluating player talent. Teams are shifting away from metrics such as batting average and slugging percentage toward advanced metrics, and may be rewarding true value more frequently as a result.

TABLE 4— SALARY COMPARED IN 2000, 2008 AND 2016 (IN REAL DOLLARS)

	Dependent variable:					
	2000 *OLS*	Robust SE *coefficient test*	2008 *OLS*	Robust SE *coefficient test*	2016 *OLS*	Robust SE *coefficient test*
	(1)	(2)	(3)	(4)	(5)	(6)
Weighted Runs Above Average	14.439***	14.439***	34.928***	34.928***	39.590***	39.590***
	(3.814)	(3.344)	(7.598)	(8.198)	(8.968)	(10.949)
Age	239.942	239.942	1,040.473**	1,040.473*	1,393.848**	1,393.848*
	(239.622)	(238.543)	(376.019)	(494.047)	(458.114)	(611.902)
Age^2	-4.654	-4.654	-16.791**	-16.791*	-24.864***	-24.864*
	(3.883)	(3.986)	(6.009)	(8.283)	(7.360)	(10.475)
Last Year's Salary	1.017***	1.017***	0.930***	0.930***	1.013***	1.013***
	(0.028)	(0.031)	(0.029)	(0.035)	(0.031)	(0.033)
Constant	-2,554.766	-2,554.766	-15,064.280**	-15,064.280*	-18,546.810**	-18,546.810*
	(3,646.497)	(3,526.115)	(5,798.950)	(7,259.972)	(7,058.635)	(8,910.410)
Observations	340		290		269	
R^2	0.865		0.866		0.867	
Adjusted R^2	0.863		0.864		0.865	
Residual Std. Error	1,277.243 (df = 335)		1,890.428 (df = 285)		2,182.928 (df = 264)	
F Statistic	536.022*** (df = 4; 335)		459.139*** (df = 4; 285)		431.873*** (df = 4; 264)	

Note: *p<0.05; **p<0.01; ***p<0.001

Last year's salary is also statistically significant to the 0.001 level in all three regressions. In fact, it contributes to R-squared more than any other variable as shown in (Table 5.) In an individual regression on a sample size of 269, last year's salary is associated with an R-Squared value of 0.844. An individual regression on wRAA with a sample size of 355 has a 0.085 R-Squared value. Both are statistically significant at the 0.001 level. Individually, with a sample size of 355, age has an R-Squared value of 0.273 at a significance level of .01. This suggests that, although it is significant, performance is not the main factor in this regression model. In most cases, the effect of last years salary eclipses the effects of wRAA.

TABLE 5— INDEPENDENT VARIABLES INDIVIDUALLY COMPARED (IN REAL DOLLARS)

	Dependent variable:					
	Weighted Runs Above	Robust SE	Last Year's Salary	Robust SE	Age	Robust SE
	OLS	*coefficient test*	*OLS*	*coefficient test*	*OLS*	*coefficient test*
	(1)	(2)	(3)	(4)	(5)	(6)
Weighted Runs Above Average	113.450***	113.450***				
	(19.863)	(22.344)				
Last Year's Salary			0.984***	0.984***		
			(0.026)	(0.026)		
Age					1,868.515*	1,868.515
					(805.393)	(1,206.174)
Age^2					-18.315	-18.315
					(13.367)	(20.866)
Constant	4,340.925***	4,340.925***	884.866***	884.866***	-33,914.700**	-33,914.700*
	(295.899)	(281.170)	(196.488)	(118.806)	(12,001.420)	(17,176.190)
Observations	355		269		355	
R^2	0.085		0.845		0.277	
Adjusted R^2	0.082		0.844		0.273	
Residual Std. Error	5,434.941 (df = 353)		2,347.809 (df = 267)		4,837.871 (df = 352)	
F Statistic	32.622*** (df = 1; 353)		1,454.597*** (df = 1; 267)		67.340*** (df = 2; 352)	
Note:					*p<0.05; ** p<0.01; ***p<0.001	

5.2 Performance and Salary by Age

The pre-free-agency market for baseball players mainly affects younger players, generally those under the age of 30. I wanted to study whether there was a difference between the effect of wRAA on real salary on younger and older players. I broke up my 2016 dataset into 141 players older than 29 and 128 players younger than 30. I then performed a regression analysis on both of the subsets. The results are presented in (table 6.) The regression on players 30 and older produced an R-Squared value of 0.847 and the regression on players younger than 30 produced an R-Squared value of 0.838.

For players who are 30 years or older, wRAA is statistically significant at the 0.01 level. For players who are younger than 30, wRAA does not have a statistically significant effect. This discrepancy could partially be explained by the restrictions on the market for younger players. Younger players are

TABLE 6— COMPARING SUBSETS OF AGE (2016) (IN REAL DOLLARS)

	Dependent variable:			
	30 Years and Older *OLS*	Robust SE *coefficient test*	29 Years and Younger *OLS*	Robust SE *coefficient test*
	(1)	(2)	(3)	(4)
Weighted Runs Above Average	48.971**	48.971**	21.592*	21.592
	(14.757)	(15.008)	(9.060)	(15.838)
Age	-3,467.468*	-3,467.468*	2,929.824	2,929.824
	(1,650.601)	(1,632.772)	(2,282.723)	(2,167.496)
Age^2	45.053	45.053	-55.336	-55.336
	(23.942)	(23.328)	(42.965)	(41.705)
Last Year's Salary:	0.996***	0.996***	1.225***	1.225***
	(0.040)	(0.030)	(0.054)	(0.116)
Constant	65,400.660*	65,400.660*	-38,181.530	-38,181.530
	(28,219.560)	(28,351.280)	(30,243.030)	(28,101.410)
Observations	141		128	
R^2	0.851		0.843	
Adjusted R^2	0.847		0.838	
Residual Std. Error	2,496.364 (df = 136)		1,544.297 (df = 123)	
F Statistic	194.762*** (df = 4; 136)		164.967*** (df = 4; 123)	

Note: *p<0.05; **p<0.01; ***p<0.001

not afforded the opportunity to earn much more than the league minimum. Even very valuable players earn relatively little per season at this stage in their career. Conversely, many players over the age of 30 have the opportunity to enter free agency, where they can be rewarded for their performance. The age variable is statistically significant at the 0.05 level in the subset of players who are 30 years and older, it is not statistically significant in the younger than 30 subset. This difference could be explained by the fact that the lower than 30 group covers only prime years of production. Younger players with lower production may have enough potential to persuade teams to offer them the same relatively low contract that they are paying their stars at the same age. The 30 and older group contains players who are aging out of their prime. Therefore, teams may be less inclined to give an aging player a lucrative contract than they are to give a player in their low 30s. Last year's salary is statistically significant in both models at the 0.001 level. It is the only significant independent variable in the less than 30 group. This reinforces the general trend that younger players have less salary variation, and flexibility.

5.3 Pre and Post Moneyball

Moneyball (Lewis, 2003) is the book that popularized sabermetrics (advanced baseball statistics) prior to its publication sabermetrics remained unknown to most fans, and the knowledge of only a few Major League Teams. I was interested in looking at the year before and after the release of this famous book. I was interested in capturing the change in the way teams were looking at player evaluation in the era in which the most famous book about the topic was written. The regression for 2002, on a sample size of 303, produced an R-Squared value of 0.868. The regression in 2004, on 299 observations, produced an R-Squared value of 0.794. In both regressions, wRAA is significant at the 0.001 level. There is a large increase in the coefficient between the two periods. It is possible that teams were beginning to change their contract practices to maximize batting value. Last year's

salary is also statistically significant at the 0.001 level. Age is only statistically significant in 2004 at the 0.01 level. It is a trend in all of these regression comparisons that over time, the effect of age on real salary has increased. This could be due to a number of factors, including an increased focus on defensive value (which older players seem to lose before batting ability), a decrease in team loyalty, or as a result of the end of the steroid era (which may have explained longevity in the first place.

TABLE 7— BEFORE AND AFTER THE RELEASE OF MONEYBALL (IN REAL DOLLARS)

	Dependent variable:			
	2002 (Before MoneyBall released)	Robust SE	2004 (After MoneyBall Released)	Robust SE
	OLS	*coefficient test*	*OLS*	*coefficient test*
	(1)	(2)	(3)	(4)
Weighted Runs Above Average	35.634***	35.634***	46.523***	46.523***
	(5.341)	(7.986)	(7.875)	(3.344)
Age	760.798*	760.798	768.676*	768.676**
	(320.253)	(453.352)	(388.880)	(238.543)
Age^2	-13.255**	-13.255	-14.126*	-14.126***
	(5.086)	(7.454)	(6.078)	(3.986)
Last Year's Salary:	0.902***	0.902***	0.836***	0.836***
	(0.027)	(0.039)	(0.032)	(0.031)
Constant	-9,903.956*	-9,903.956	-9,615.770	9,615.770**
	(4,976.742)	(6,809.296)	(6,131.436)	(3,526.115)
Observations	303		299	
R^2	0.870		0.797	
Adjusted R^2	0.868		0.794	
Residual Std. Error	1,556.422 (df = 298)		2,230.656 (df = 294)	
F Statistic	499.454*** (df = 4; 298)		287.996*** (df = 4; 294)	
Note:				*p<0.05; **p<0.01; ***p<0.001

6. Conclusions

6.1 What Does the Model Tell Us?

The regression models presented today almost all help to conclude that weighted runs above average is associated with a change in the real salary of a player. It seems that this effect is more pronounced in older players than younger players. Every model helps to conclude that previous salaries are associated with future salaries. This is not a novel discovery; we already know that contracts last longer than a year. The constant values in all of the regressions are not meaningful alone. Most of the constants are negative and no player is paying a

team to play for them. Intuitively, age is a factor in almost every one of the regression models. This is not surprising either, since we expect there to be diminishing returns on performance with age. Ultimately, I was surprised at the relatively low coefficients on wRAA. I hypothesized that contracts would be largely correlated with batting performance.

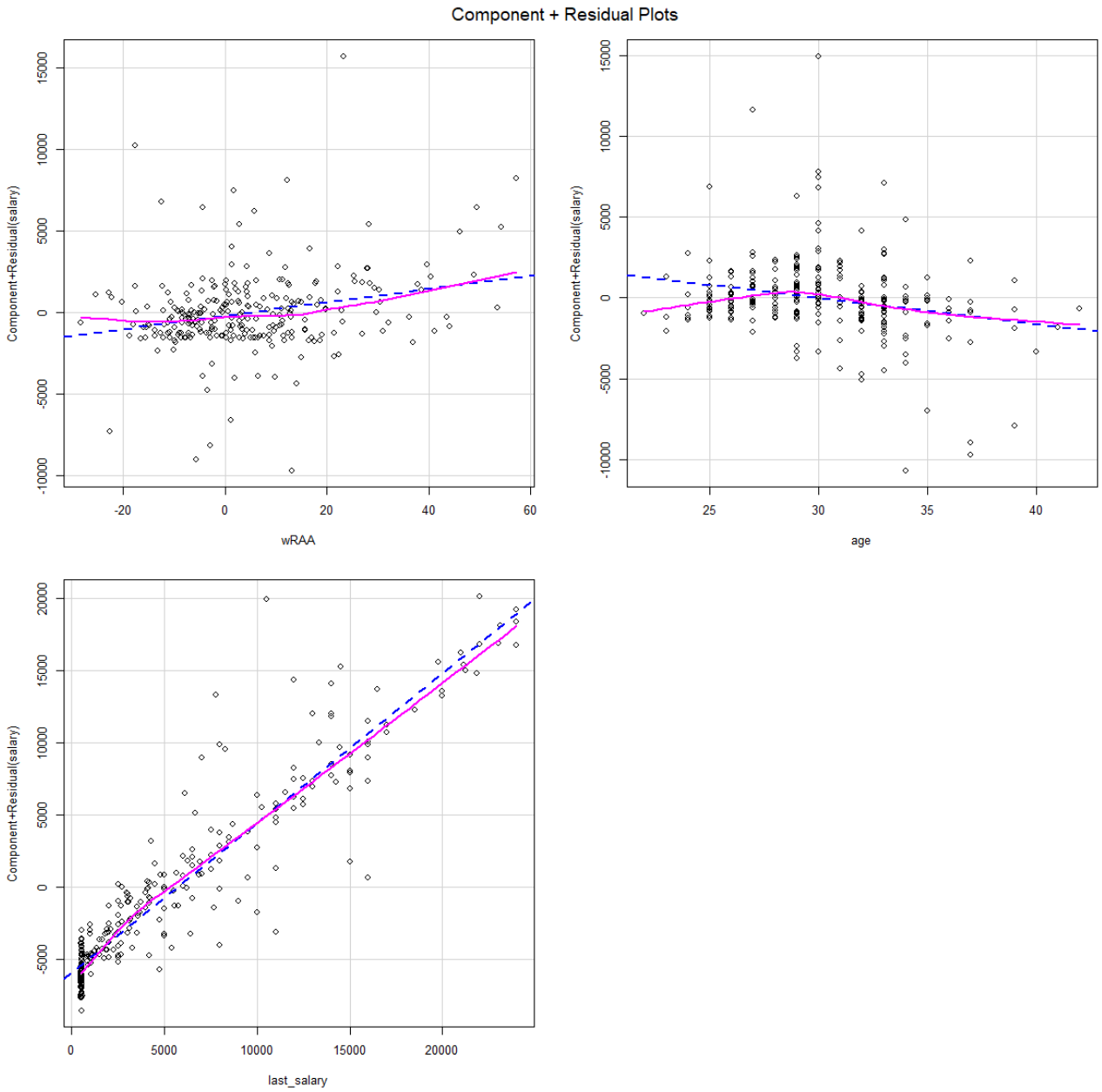
6.2 What Is Missing?

The model presented in this paper ultimately suffers from a dearth of information, and consequently, missing variable bias. Teams have always known that batting was valuable, it is the most visible, and isolated, aspect of the game. Modern metrics have made much deeper strides in evaluation of pitching and defense. I suspect that if I could source information on defensive value as well as modern pitching statistics, I would be able to control for variables that have expected effects on real salary. I believe that controlling for these variables would explain some of the predictive value of last year's salary.

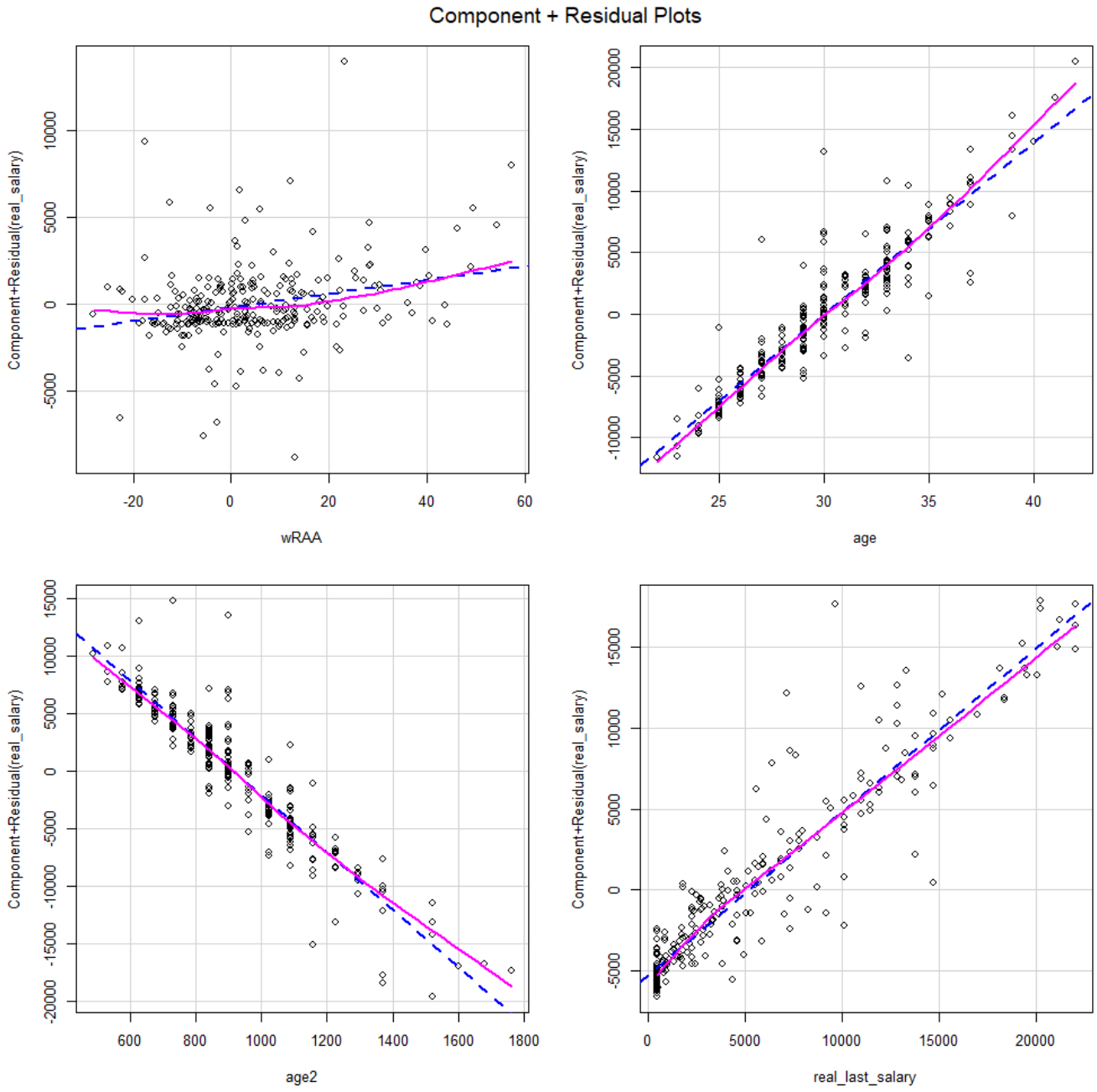
Regressing each year individually was not optimal, by restricting my regressions to one year at a time I sacrificed some of the predictive value of a large dataset over many years for simplicity's sake. I would be interested in researching these issues further using the entire dataset as panel data, or with an entirely different statistical method such as an instrumental variable model.

I recognize the shortcomings of the model that I have presented, but I believe that it presents some meaningful, and useful, inferences that may serve as encouragement to perform more meaningful analysis.

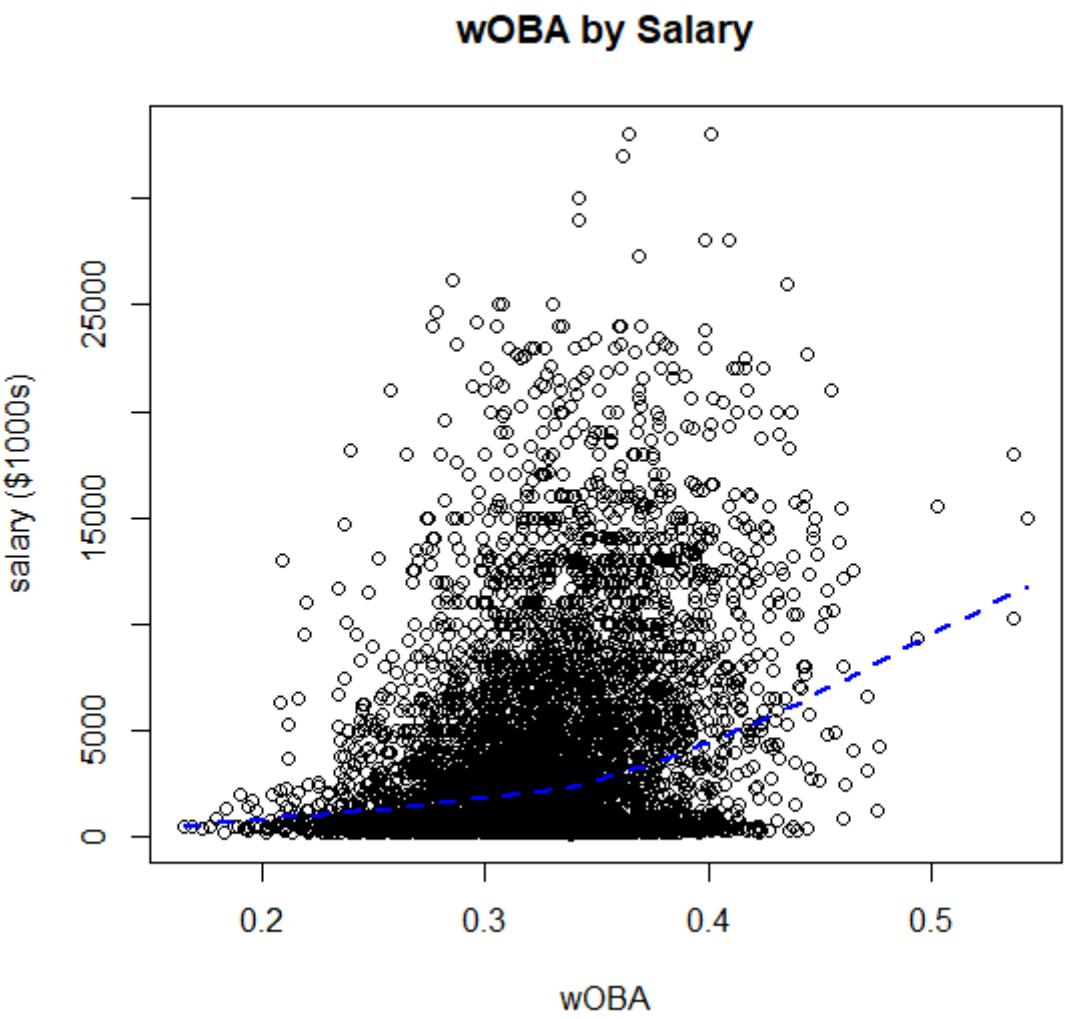
Appendix



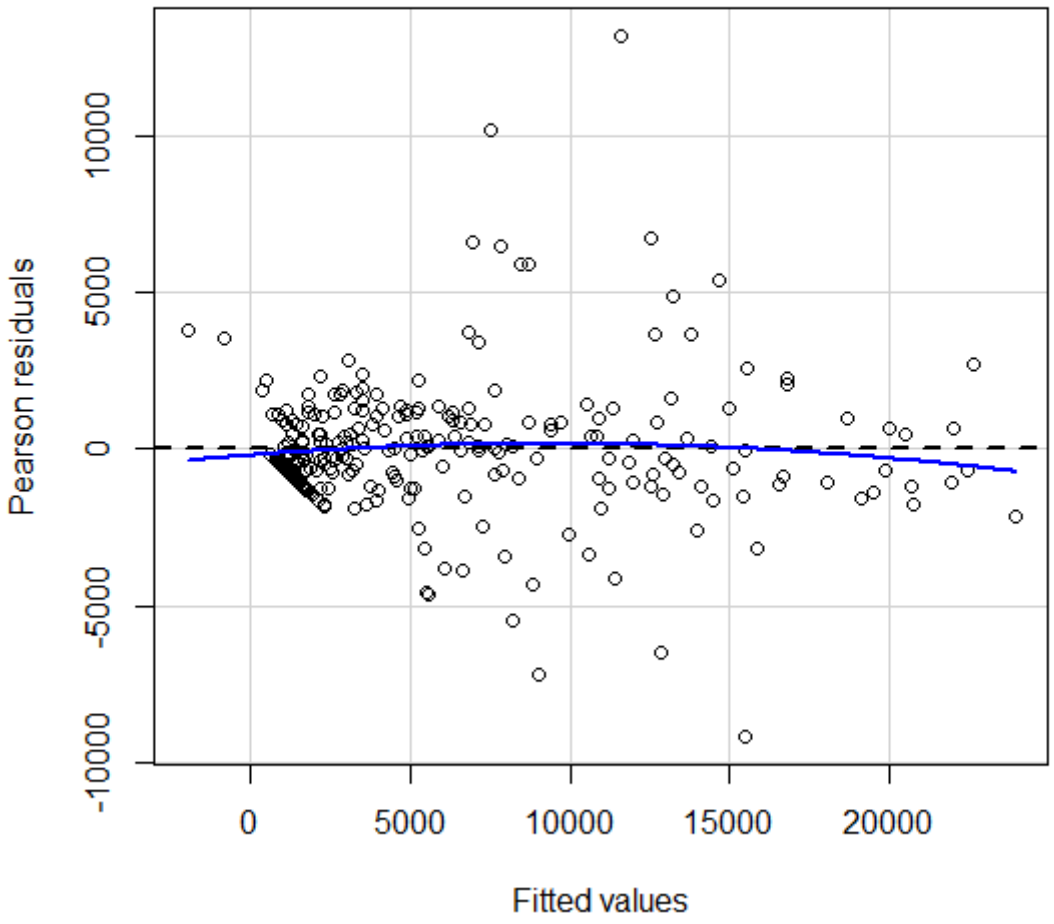
(Figure 4. CR Plots without quadratic age)



(Figure 5. CR Plots with quadratic age)



(Figure 6. wOBA by Salary)



(Figure 7. Residual Plot for Dependent variable: real salary)

References

[1] Bennett, J. M., & Flueck, J. A. (1983). An Evaluation of Major League Baseball Offensive Performance Models. *The American Statistician*, *37*(1), 76–82. https://doi.org/10.1080/00031305.1983.10483093

[2] Schall, T., & Smith, G. (2000). Do Baseball Players Regress toward the Mean? *The American Statistician*, *54*(4), 231–235. https://doi.org/10.1080/00031305.2000.10474553

[5] O'Neill, H. M. (2013). Do Major League Baseball Hitters Engage in Opportunistic Behavior? *International Advances in Economic Research*, *19*(3), 215–232. https://doi.org/10.1007/s11294-013-9419-y

[6] Tymkovich, J. (2012). A Study of Minor League Baseball Prospects and Their Expected Future Value. *CMC Senior Theses*. Retrieved from https://scholarship.claremont.edu/cmc_theses/442

[3] Bradbury, J. C. (2007). Does the Baseball Labor Market Properly Value Pitchers? *Journal of Sports Economics*, *8*(6), 616–632. https://doi.org/10.1177/1527002506296366

[4] Beneventano, P., Berger, P., & Weinberg, B. (2019). *Predicting Run Production and Run Prevention in Baseball: The Impact of Sabermetrics*.

[7] wOBA | FanGraphs Baseball. Retrieved May 29, 2019, from https://www.fangraphs.com/guts.aspx?type=cn

[8] World Bank Group - International Development, Poverty, & Sustainability. Retrieved May 31, 2019, from World Bank website: http://www.worldbank.org/

[9] Collective Bargaining Agreement. (n.d.). Retrieved May 31, 2019, from MLBPlayers.com website: http://www.mlbplayers.com/ViewArticle.dbml?ATCLID=211078089&DB_OEM_ID=34000

[10] Schulz, R., Musa, D., Staszewski, J., & Siegler, R. S. (1994). The relationship between age and major league baseball performance: Implications for development. *Psychology and Aging*, *9*(2), 274–286. https://doi.org/10.1037/0882-7974.9.2.274

[11] Lewis, M. (2003). Moneyball: The art of winning an unfair game. New York: W.W. Norton.