2020

A Replication and Cross-Validation of Hakes and Sauer's "An Economic Evaluation of the Moneyball Hypothesis"

Joshua Ross Audette
University of Vermont

A Replication and Cross-Validation of Hakes and Sauer's "An Economic Evaluation of the *Moneyball* Hypothesis"

Josh Audette*

May 2020

This paper replicates and cross-validates the econometric models presented by Hakes and Sauer in "An Economic Evaluation of the *Moneyball* Hypothesis." The authors suggest an inefficiency in the major league baseball player market leading up to the publishing of Michael Lewis's *Moneyball*. Through the usage of basic econometric methods, Hakes and Sauer (2006) determined that while on-base and slugging percentage were improperly valued prior to the book's release in 2003, the inefficiency was corrected soon after. This study is successful in replicating the results produced by Hakes and Sauer, as well as cross-validating their findings with an agent-based model. However, after extending their work with new and current data, the external validity of their model is brought into question. Although major league baseball teams properly valued on-base and slugging percentage following the release of *Moneyball*, this correction no longer holds when examining data beyond the 2004 season.

1 Introduction

It is important to first define the phrase "*Moneyball* hypothesis" as is used in the title of Hakes and Sauer (2006). This is in reference to Michael Lewis's groundbreaking book *Moneyball*, published in 2003. The book follows the Oakland Athletics, who were successful in gaining an edge on opposing teams through their utilization of sabermetrics in the player valuation process. Specifically, the Athletics developed the theory that teams in major league baseball (MLB) were overvaluing slugging percentage (SLG) and undervaluing on-base percentage (OBP) in terms of their contribution to wins (Lewis, 2003). By putting this theory into practice, and acquiring players with a high OBP, Oakland became one of the best teams in the American League. After Lewis (2003) brought the theory behind the Athletics' success into the public eye, Hakes and Sauer sought

*Version 1.1. May 2020. Contact jaudette08@gmail.com. University of Vermont, Economics. I wish to thank William Gibson for his advising and guidance, Richard Paulsen for help with data collection, and Kevin Issadore for his previous work (Issadore, 2009) and inspiration. All remaining errors are my own.

to corroborate the idea through basic econometric procedures. This paper successfully replicates the procedures used by Hakes and Sauer which confirmed a market inefficiency in the MLB leading up to the 2004 MLB season. Furthermore, this paper builds on Hakes and Sauer's study through the usage of up-to-date data. In doing so, the lasting effect that the "*Moneyball* hypothesis" had on the MLB player market is scrutinized.

The remainder of the paper is organized as follows. Section 2 presents the replications of Hakes and Sauer's regression determining the effect of OBP and SLG on winning percentage, as well as introduces the details and results of the agent-based model. Section 3 displays and discusses the replications of Hakes and Sauer's labor market regressions. Section 4 examines the internal and external validity of Hakes and Sauer's models. Section 5 presents conclusions. Section 6 displays the pseudo-code of the agent-based model. Section 7 is the appendix.

2 How OBP and SLG Effect Winning Percentage

2.1 Relevant offensive statistics

The goal of a MLB team is to win games by scoring the most possible runs over the course of nine innings. For each inning, the batting team is allowed three outs; equating to twenty-seven outs per game per team.[1] It is obvious that the team who is the best at avoiding outs will score the most runs and thus win the most games.

There have been multiple metrics, some dating back to the early history of the sport, that attempt to quantify a hitter's contribution to both scoring runs and winning games. Some of these metrics are more efficient than others. Batting average (BA), historically the most commonly used measure of batting proficiency, is the ratio of total hits to total at-bats. However, it does not take into account walks or hit by pitch. The metric also weighs a home run equal to that of a single, even though it is clear that one is more productive than the other.

Slugging percentage, S_lg, remedies this issue by awarding different weights to different types of hits. It can be calculated via the following equation:

$$S_lg = \frac{1_b + 2(2_b) + 3(3_b) + 4(H_r)}{A_b} \tag{1}$$

where 1_b is singles, 2_b is doubles, 3_b is triples, H_r is home runs, and A_B is at-bats. This statistic rises as a player hits more extra-base hits, and thus gives less value to non-power-hitting players. However, it completely ignores productive non-out-producing outcomes such as walks and hit by pitch. These are properly

[1] This is true unless the home is winning going into the second half of the ninth inning, in which case they only play twenty-four of those outs as the last half-inning becomes meaningless.

accounted for in the OBP metric. Despite a single and a walk both resulting in the batter ending up on first base, BA and SLG fail to account for its value. OBP does, and can be seen expressed below; where OBP is denoted as O_{bp}, hits as H, walks as B_b, hit by pitch as H_{bp}, at-bats as A_b, and sacrifice hits as S_h:

$$O_{bp} = \frac{H + B_b + H_{bp}}{A_b + B_b + H_{bp} + S_h} \tag{2}$$

Sacrifice hits are factored into the denominator because they are also not naturally included in the at-bat statistic. A sacrifice hit is either a sacrifice fly or sacrifice bunt which results in an out, but also scores a run. A skilled hitter can execute either of these at will if the situation calls for it. Because OBP is a measure of a batter's ability to reach base safely and avoid outs, it is important to include all events in which an out is recorded, including sacrifices.

2.2 OBP and SLG's effect on wins

Hakes and Sauer (2006) sought to determine whether OBP and SLG were properly valued in the MLB player market prior to and soon after the release of Michael Lewis's *Moneyball*. In determining this, it is critical to first properly estimate both statistics' contribution to wins.

Table 1: Team OBP and SLG Data Summary

Variable	Obs	Mean	Std. Dev.	Min	Max
OBP[1]	630	0.328	0.015	0.292	0.373
OBP Against[2]	630	0.328	0.016	0.282	0.384
SLG[3]	630	0.417	0.027	0.335	0.495
SLG Against[4]	630	0.417	0.028	0.341	0.499
WPCT[5]	630	0.500	0.073	0.265	0.716

Notes: 1. Team on-base percentage.
2. Team on-base percentage against.
3. Team slugging percentage.
4. Team slugging percentage against.
5. Team winning percentage.
Source: Author's computations based on data gathered from Baseball-Reference (2019).

Ordinary least squares (OLS) is used to estimate the partial effect of team OBP and team SLG on team winning percentage (WPCT). WPCT is simply

Table 2: The Effect of OBP and SLG on Winning: Hakes and Sauer (2006)

WPCT	(1)	(2)	(3)	(4)
OBP	3.294		2.141	2.032
	(0.092)		(0.296)	(0.183)
OBP Against	-3.317		-1.892	-2.032^R
	(0.088)		(0.126)	
SLG		1.731	0.802	0.900
		(0.122)	(0.149)	(0.106)
SLG Against		-1.999	-1.005	-0.900^R
		(0.122)	(0.152)	
Constant	0.508	0.612	0.502	0.500
	(0.114)	(0.073)	(0.099)	(0.005)
R^2	0.825	0.787	0.885	0.884
Observations	150	150	150	150

Standard errors in parentheses. All p-values significant at the one percent level.
Notes: Dependent variable is winning percentage. Superscript "R" denotes the coefficient was set equal to its counterpart. Variables are the same as table 1.
Source: Hakes and Sauer (2006).

the amount of games a team wins divided by 162—the amount of games that are played in an MLB season. The variables of *OBP against* and *SLG against* are also included in the regression analysis. These are average levels of OBP and SLG that a team faces throughout a season. These variables are added in order to give a more complete picture of the true effect of OBP and SLG. It allows the model to take into account the contribution of the team's own OBP and SLG, as well as that of their opponent's on winning percentage. This eliminates some of the contribution that a proficient pitching staff has on WPCT by factoring the success of opposing hitters into the equation. If a team has a first-rate pitching staff, then on-base percentage against and slugging percentage against would be very low. Also, it eliminates the effect that the quality of hitters on the opposing team has on WPCT.

However, as is explained in Hakes and Sauer (2006), the coefficients on OBP and OBP against, as well as SLG and SLG against, will be very similar since the model is examining a two-sided symmetric game.

Table 3: The Effect of OBP and SLG on Winning: 1999 - 2004

WPCT	(1)	(2)	(3)	(4)
OBP	3.296		2.113	2.034
	(0.184)		(0.229)	(0.162)
OBP Against	-3.325		-1.922	-2.034^R
	(0.195)		(0.240)	
SLG		1.749	0.964	0.902
		(0.113)	(0.139)	(0.096)
SLG Against		-1.971	-0.994	-0.902^R
		(0.108)	(0.144)	
Constant	0.509	0.595	0.511	0.500
	(0.097)	(0.061)	(0.075)	(0.001)
R^2	0.824	0.785	0.884	0.884
Observations	150	150	150	150

Robust standard errors in parentheses. All p-values significant at the one percent level.
Notes: Dependent variable is winning percentage. Superscript "R" denotes the coefficient was set equal to its counterpart. Variables are the same as table 1.
Source: Author's computations based on data gathered from Baseball-Reference (2019).

Table 4: The Effect of OBP and SLG on Winning: 1999 - 2019

WPCT	(1)	(2)	(3)	(4)
OBP	3.093		1.746	1.825
	(0.092)		(0.128)	(0.099)
OBP Against	-3.193		-1.890	-1.852^R
	(0.088)		(0.126)	
SLG		1.760	0.964	0.955
		(0.053)	(0.072)	(0.058)
SLG Against		-1.913	-0.945	-0.955^R
		(0.049)	(0.073)	
Constant	0.533	0.564	0.539	0.500
	(0.039)	(0.026)	(0.033)	(0.001)
R^2	0.766	0.749	0.832	0.831
Observations	630	630	630	630

Robust standard errors in parentheses. All p-values significant at the one percent level.
Notes: Dependent variable is winning percentage. Superscript "R" denotes the coefficient was set equal to its counterpart. Variables are the same as table 1.
Source: Same as table 1.

2.3 Can the regressions of WPCT on OBP and SLG produced by Hakes and Sauer (2006) be replicated?

Table 1 displays the summary statistics for performance data gathered from the 1999 - 2019 MLB seasons. Although Hakes and Sauer (2006) only used data from the 1999 - 2004 seasons, current data is utilized in order to extend the reach of their study. In order to first replicate the findings of Hakes and Sauer (2006), four regressions were run on the data from 1999 - 2004. These estimates were successful in replicating the results of Hakes and Sauer (2006), and can be seen in table 3. Column one displays a regression of WPCT on OBP and OBP against. The estimate from this regression shows that the R^2, 0.824, is almost identical to that of Hakes and Sauer (2006), 0.825, which can be seen in table 2. The values only differ by 0.001—less than one percent. This is also the case for the following three columns, all of which present R^2 values that only stray from Hakes and Sauer's results by 0.002 at the most.

Table 5: Correlation Matrix

Variable	OBP	OBP Against	SLG	SLG Against	WPCT
OBP	1.000				
OBP Against	0.159	1.000			
SLG	0.749	0.124	1.000		
SLG Against	0.111	0.812	0.220	1.000	
WPCT	0.512	-0.620	0.487	-0.591	1.000

Notes: Variables are the same as table 1.
Source: Same as table 1.

Although the coefficients are not exactly equivalent, the estimates seem to successfully mimic the results that were achieved by Hakes and Sauer (2006). The data source originally used by Hakes and Sauer (2006) is no longer available, which most likely is the cause of the small difference in the regression results. The fourth column of table 3 analyzes the "*Moneyball* hypothesis" (Lewis, 2003). To achieve this, the *OBP against* and *SLG against* variables are restricted to their counterparts. The resulting linear regression equation is as follows.

$$W_{pct} = \beta_0 + \beta_1 \times (OBP - OBP_a) + \beta_2 \times (SLG - SLG_a) \quad (3)$$

Where W_{pct} is winning percentage, OBP is on-base percentage, OBP_a is on-base percentage against, SLG is slugging percentage, and SLG_a is slugging percentage against. By taking the difference between OBP and OBP against, as well as SLG and SLG against, the regression is able to properly estimate the partial effect of OBP and SLG on WPCT. The equation is thus showing the impact of being one point better or worse than your opponent in OBP or SLG. As can be seen in column four of table 3, the data suggests that, in comparison to SLG, OBP has over two-times the effect on winning percentage. In terms of wins, a one-point increase in OBP is far more valuable than a one-point increase in SLG. The coefficients on OBP and SLG both only differ from those of Hakes and Sauer (2006) by 0.002.

However, it is worth mentioning that, given the nature of how the two statistics are calculated, OBP and SLG are correlated at a high level. As can be seen in both equation 1 and equation 2, most outcomes that do not result in an out will increase both OBP and SLG. This is indicative of imperfect multicollinearity; meaning that the two independent variables are highly correlated. This relationship of the two independent variables can be seen clearly in in table 5—a correlation matrix of all variables found in the regression analysis. OBP and SLG have a correlation coefficient of 0.749 out of 1.000.

The high correlation between OBP and SLG suggests that it *may* be difficult to produce a fully precise estimate of the partial effect of the two independent variables on the dependent variable. However, with each coefficient in the regression significant at the one percent level, this correlation does not pose a major threat to the precision of the estimate.

Table 4 shows the continuation of Hakes and Sauer's regression using up-to-date data spanning from 1999 - 2019. The results are similar to that of tables 3 and 2. The R^2 values decrease slightly overall, as do the coefficients on OBP. The coefficients on SLG are marginally higher than those found in Hakes and Sauer (2006), which can possibly be attributed to the overall increase in home runs league-wide in the MLB.[2] Despite this, the effect of OBP on winning percentage still remains almost double that of SLG over the course of the 1999 - 2019 MLB seasons.

Both regressions are successful in replicating the results of Hakes and Sauer (2006). The coefficients found in both replication tables show the expected signs and similar values to those of Hakes and Sauer (2006), as well as significance at the one percent level.

2.4 Cross-validating Hakes and Sauer (2006) with an agent-based model

To further confirm the findings of Hakes and Sauer (2006), an agent-based model is employed. An agent-based model (ABM) is a computational model which allows for the simulation of actions and outcomes using autonomous "agents." An ABM produces randomized control trials which are not biased by unobservable variables and factors. Within a baseball game this would include factors such as weather, variation in ballparks, luck, nervousness of players, etc. The usage of an ABM in this instance will act as a means of cross-validation for the regression of Hakes and Sauer (2006) which aims to determine OBP and SLG's effect on WPCT. Verification using an ABM, coupled with the successful replication and extension seen in the section above, will further solidify the findings pertaining to the effect of OBP and SLG on WPCT.

The ABM utilized in this paper is an in-depth simulation of baseball games and seasons. Each team is endowed with a set of batting statistics (team OBP and team SLG) which represent the average for the players on their team. These statistics last for the entire 162-game season, and at the end of the season the amount of wins, losses, ties, level of OBP, and level of SLG is recorded. At the beginning of a new season, each team is given a new set of random batting statistics. This way the changes in a team's winning percentage at the end of the season can be attributed to the fashion in which their OBP and SLG has changed. The following sections detail how, exactly, the model works, as well as the results.

[2] Average MLB home runs per season has increased 22 percent from 1999 - 2019 (Baseball-Reference, 2019).

2.4.1 Team creation and endowment

At the beginning of each simulated season, thirty teams are created to embody all thirty organizations in the MLB.

The maximum and minimum team OBP in the MLB from the 1999 - 2019 seasons, as can be seen in table 1, was 0.373 and 0.292, respectively (Baseball-Reference, 2019). The maximum and minimum team SLG in the MLB for the 1999 - 2019 seasons was 0.335 and 0.495, respectively (Baseball-Reference, 2019). Each team's randomly assigned values of OBP and SLG are between these MLB minimum and maximum ranges. Each team's assigned OBP is then split into the probability that the outcome of the at-bat will be a single, double, triple, home run, walk, or hit by pitch. Average data from the 1999 - 2019 MLB seasons is again used to find the average rates at which each of these outcomes take place. For all outcomes in which a player got on base in the MLB during this time frame, seventy-one percent of the them were either a single, walk, or hit by pitch (Baseball-Reference, 2019).[3] Sixteen percent of the time the hit was a double, 0.9 percent of the time a triple, and twelve percent of the time a home run (Baseball-Reference, 2019). Using these percentages, a team's OBP can be split up so each component represents a different outcome. For instance, if a team is randomly endowed with a 0.300 OBP, then 0.213 of that will be either singles, walks, or hit by pitch, 0.048 will be doubles, 0.0027 triples, and 0.036 home runs.

All thirty teams thus have a unique probability that, for each at-bat, the outcome will be a single, double, triple, home run, walk, or hit by pitch. Sixty players are then created and move randomly to a team so that each team has two players.[4] These players adopt the endowed batting statistics of the team to which they have moved to. Each team, as well as its players, are now ready to engage in the game-playing sequence.

2.4.2 The game

A single baseball game is made up of nine scheduled innings, and each inning separated into two half-innings. For each half-inning, one team hits and one team pitches and plays defense. The hitting team is allowed three outs, at which point the half-inning ends and the other team comes up to hit. After the nine innings have been completed, whichever team has accrued the most runs wins the game.

At the beginning of each season within the model, all thirty teams randomly match-up against another team and are randomly assigned as either the home or visiting team. Following this, the visiting team begins the playing sequence, as the visiting team always hits first in a baseball game. One of the two players on the visiting team becomes the batter and is given a "pitch." This "pitch"

[3]Having effectively the same outcome on a game, singles, walks, and hit by pitch are combined into one set of probabilities in this model.

[4]Two players are used per team for ease of use within the model. Given that each player on the team has the same batting statistics, the amount of players per team has no bearing on the outcome of the simulation.

is simply a random decimal between zero and one. If the decimal is greater than that of the player's endowed OBP, then the player is out. For instance, if a player has an OBP of 0.297 and the random "pitch" is 0.642, then that player is out and the other player on the team comes up to hit.[5] If, for the same player previously mentioned, the "pitch" is 0.241, then the player gets a hit and gets on base. The type of hit is determined by the average MLB probabilities mentioned in the preceding section. The player who was awarded the hit now occupies the base corresponding to the type of hit. If it is a single, walk, or hit by pitch, the player is on first base; if it is a double the player is on second base; if it is a triple the player is on third base; and if it is a home run there is no one on base and a run is scored. The two players on the team switch off every at-bat no matter the outcome.

If there is already someone on-base when a player comes up to hit, then there are multiple outcomes which can take place. With a player, or players, on-base, a single will move everyone up one base and put the hitter on first base. Thus if there is a player on third base, they will score. However, as is the case in real MLB baseball games, a player has a reasonable chance to score if they are on second base and a single is hit. According to the MLB average from 1999 - 2019, the probability that a player will score in this scenario is forty percent (MLB.com, 2019). Therefore, if there is a player on second base and the outcome of the at-bat is a single, forty percent of the time they will score and sixty percent of the time they will simply advance to third base. A double will move all players up two bases and put the hitter on second base. Any runners occupying second or third base will score. If there is a player on first base when a double is hit, then there is a fifteen percent chance that runner will score (MLB.com, 2019). Therefore, in the model, when a double is hit and a runner is on second base, fifteen percent of the time the player will score a run, and eighty-five percent of the time they will advance two bases to third base. A triple will score all players who are on base and put the hitter on third base. A home run will clear the bases and score all players who are on base, as well as the hitter themselves. If there is a player on first base with less than two outs, and the outcome of the at-bat is an out, then there is a fourteen percent chance that the out will result in a double-play (MLB.com, 2019). This means that, with less than two outs and a player on first base, if the at-bat results in an out, then fourteen percent of the time two outs will be recorded and the player on first base will be cleared. Eighty-six percent of the time one out is recorded and no runners advance.

After three outs are recorded, the half-inning is over and the runs are totaled to represent the current score of the visiting team. The home team then comes up to hit, repeating the same sequence as the visiting team until three outs are recorded and the first inning has concluded. After nine innings, the game is over and the total amount of runs from each team is tallied. The team with more runs is awarded with a win and the other with a loss. If the game ends

[5] The type of out (strikeout, ground-out, fly-out, etc.,) is not relevant to the conclusions drawn in this paper or the simulated game.

and the score is tied, then both teams are awarded a tie.

2.4.3 The season

Following the completion of a game, all thirty teams subsequently match-up at random with a different team to play their next game. This is not true to how match-ups work in the MLB, as teams mainly play those within their league and division. However, the quality of the teams within a league and division is for the most part random in the MLB (at least in theory), and should not have an effect on the accuracy of this study. This happens until all thirty teams have played 162 games, at which point the season ends. The total wins, losses, and ties are recorded for each team, as well as their endowed team OBP and team SLG. One season is now over, and the teams and players reset. The next season begins, and the teams are endowed with a new set of OBP and HRP, with players moving to different teams to adopt these new batting statistics. Twenty-one seasons are simulated to reflect the same amount of observations as table 4, with the season totals recorded at the conclusion of each season.

2.4.4 Does the agent-based model validate the results of Hakes and Sauer (2006)?

Using the data from twenty-one simulated seasons, an OLS linear regression is run to analyze the significance of OBP and SLG on WPCT. Table 6 displays a general summary of the data collected throughout the simulation.

Table 6: Agent-Based Model Data Summary

Variable	Obs	Mean	Std. Dev.	Min	Max
OBP[1]	630	0.326	0.016	0.292	0.373
SLG[2]	630	0.395	0.047	0.335	0.495
WPCT[3]	630	0.509	0.081	0.309	0.722

Notes: 1. Team on-base percentage.
2. Team slugging percentage.
3. Team winning percentage.
Source: Author's computations based on data
data from the ABM.

In column one of table 7, OBP is examined individually for its impact on WPCT. It is noted that the R^2 value is 0.739, which indicates that OBP accounts for roughly 73.9 percent of the variation in WPCT among all teams in the

Table 7: The Effect of OBP and SLG on Winning: Agent-Based Model

WPCT	(1)	(2)	(3)
OBP	3.890		1.381
	(0.109)		(0.155)
SLG		1.507	1.125
		(0.034)	(0.056)
Constant	-0.759	-0.085	-0.385
	(0.035)	(0.013)	(0.035)
R^2	0.739	0.611	0.769
Observations	630	630	630

Robust standard errors in parentheses.
All p-values significant at the one percent level.
Notes: Variables are the same as table 6.
Source: Same as table 6.

model. Column two shows the regression of WPCT on SLG. The R^2 decreases in this regression to 0.611, which is consistent with the findings of Hakes and Sauer (2006). This points to OBP as a more predictive metric of WPCT when compared to SLG. These findings can also be seen expressed visually in figures 1 and 2. Although both plots show a very clear linear trend, the data cluster in the plot of OBP and WPCT is more tightly bunched when compared to the plot of SLG and WPCT. The tighter the data cluster is bunched, the better the fit of the linear regression.

Hakes and Sauer (2006) uses the independent variables *OBP against* and *SLG against* in their regression, which this model does not. This is because this model has no need to account for the skill of a pitching staff. Nothing in the model sets one team's pitchers as better than the other. Effectively, every team and every player is facing the same pitcher for every at-bat. The only factors that cause variance in WPCT is OBP and SLG. This eliminates the need for the *OBP against* and *SLG against* variables. It also eliminates other factors that slightly contribute to WPCT, such as fielding percentage.

The coefficients on OBP and SLG in columns one and two are significant, and are very close to the results of Hakes and Sauer's regression of WPCT on OBP and SLG. The coefficient on OBP in column one is 3.890, compared to 3.294 in Hakes and Sauer (2006). In column two, the coefficient on SLG is 1.507;

Figure 1: On-Base and Winning Percentage: Agent-Based Model

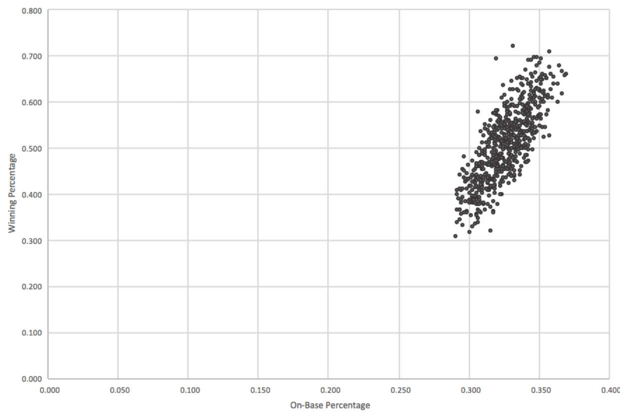
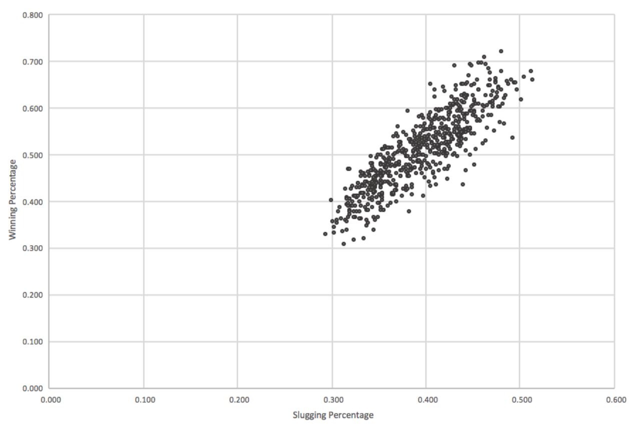


Figure 2: Slugging and Winning Percentage: Agent-Based Model



only 0.224 less than Hakes and Sauer's 1.731.

Column three of table 7 is the key indicator that the results of the ABM validates the findings of Hakes and Sauer (2006). The R^2 increases slightly to 0.769, and the coefficients on both variables decrease. However, the coefficient on OBP remains greater than that of SLG, suggesting that it has more of an effect on WPCT than SLG. This is in-line with the findings of Hakes and Sauer (2006), as well as the replications previously presented. Even though the difference between the coefficients on OBP and SLG is less than what was found in Hakes and Sauer (2006), it still shows that WPCT has a stronger positive correlation with OBP than it does with SLG. This can also be seen in the scatter plots. The slope of the cluster in figure 1 is much steeper than the slope in figure 2.

The results of the ABM are successful in cross-validating the findings of

Hakes and Sauer (2006). Along with the previous replications, this validation further strengthens the results of Hakes and Sauer's regression, which determines that OBP is more predictive of wins than SLG. By using empirical data from MLB baseball seasons, as well as the randomized control trials of the ABM, there is sufficient evidence that OBP does, in fact, contribute more to winning baseball games than SLG.

3 The MLB Labor Market's Valuation of OBP and SLG

Given the findings of Hakes and Sauer (2006) presented in the previous sections, it is apparent that OBP has more of an impact than SLG on team wins. One would expect that the labor market would value these metrics accordingly—paying more money to players who are able to consistently reach base safely, and slightly less to those who hit for extra-base hits. However, it is the hypothesis of both Lewis (2003) and Hakes and Sauer (2006) that OBP was improperly valued leading up to the 2004 MLB season.

Hakes and Sauer assess this by estimating salary equations for all non-pitchers in the MLB. This estimation can be seen in table 8, with the natural log of annual salary as the dependent variable. All statistics are gathered from year t, while log salary is retrieved from year $t + 1$. This is done because salary compensation is most often based on the player's performance from the year prior. As is done in Hakes and Sauer (2006), a 130 AB threshold is utilized to exclude players who did not play the majority of the season. *Plate appearances* (PA) is included in the regression to account for annual service time. Hakes and Sauer account for a player's salary-bargaining abilities through the binary variables *Arbitration eligible*,[6] and *Free agency*.[7] The binary variables *Catcher dummy* and *Infielder dummy* are included to control for the varying defensive importance that comes along with different positions in baseball.

Table 9 displays the replication of Hakes and Sauer's regression. Due to a lack of accessible data, this regression uses *Age* as a variable to measure bargaining power instead of *Arbitration eligible* and *Free agency*. *Age* should have a similar effect on salary, as a player's age is most likely correlated with how long they have been in the MLB, and thus their bargaining position.

Table 8: Labor Market Valuation of OBP and SLG: Hakes and Sauer (2006)

Log Salary	2000 - 2004 (1)	2000 - 2003 (2)	2004 (3)
OBP[1]	1.360	0.842	3.681
	(0.625)	(0.678)	(1.598)
SLG[2]	2.392	2.453	2.175
	(0.311)	(0.338)	(0.788)
Plate Appearances[3]	0.003	0.003	0.003
	(0.000)	(0.000)	(0.000)
Arbitration Eligible[4]	1.255	1.242	1.323
	(0.047)	(0.048)	(0.115)
Free Agency[5]	1.683	1.711	1.575
	(0.044)	(0.185)	(0.105)
Catcher Dummy[6]	0.152	0.185	0.059
	(0.056)	(0.061)	(0.133)
Infielder Dummy[7]	-0.029	-0.007	-0.100
	(0.040)	(0.044)	(0.098)
Constant	10.083	10.429	9.782
	(0.170)	(0.178)	(0.414)
R^2	0.675	0.655	0.635
Observations	1736	1402	340

Standard errors in parentheses.
Notes: 1. On-base percentage.
2. Slugging percentage.
3. Plate appearances.
4. Dummy variable for arbitration eligibility.
5. Dummy variable for free agency eligibility.
6. Dummy variable for catchers.
7. Dummy variable for infielders.
Source: Hakes and Sauer (2006).

Table 9: Labor Market Valuation of OBP and SLG: 2000 - 2004

Log Salary	2000 - 2004 (1)	2000 - 2003 (2)	2004 (3)
OBP	1.774	1.134*	4.965*
	(0.797)	(0.892)	(1.792)
SLG	2.898**	3.089**	2.116*
	(0.437)	(0.481)	(1.029)
Plate Appearances	0.003**	0.003**	0.003**
	(0.000)	(0.000)	(0.000)
Age[1]	0.116**	0.120**	0.099**
	(0.007)	(0.008)	(0.016)
Catcher Dummy	0.141	0.172	0.049
	(0.076)	(0.084)	(0.180)
Infielder Dummy	-0.080	-0.058	-0.172
	(0.054)	(0.059)	(0.139)
Constant	7.519**	7.487**	7.615*
	(0.261)	(0.288)	(0.638)
R^2	0.500	0.509	0.474
Observations	1378	1114	264

Robust standard errors in parentheses. ** $p < 0.01$, * $p < 0.05$.
Notes: 1. Age of player.
The rest of the variables are the same as table 8.
Source: Author's computations based on data gathered from SeanLahmanDatabase (2019).

Table 10: Labor Market Valuation of OBP and SLG: 2005 - 2017

Log Salary	2005 - 2008 (1)	2009 - 2012 (2)	2013 - 2017 (3)
OBP	1.495	1.110	3.566**
	(1.015)	(1.075)	(1.017)
SLG	2.883**	3.192**	2.123**
	(0.515)	(0.540)	(0.553)
Plate Appearances	0.003**	0.003**	0.003**
	(0.000)	(0.000)	(0.000)
Age	0.139**	0.149**	0.180**
	(0.008)	(0.008)	(0.009)
Catcher Dummy	0.090	0.162	-0.049
	(0.078)	(0.084)	(0.092)
Infielder Dummy	-0.026	-0.029	-0.067
	(0.061)	(0.063)	(0.064)
Constant	7.198**	7.229**	6.434**
	(0.321)	(0.309)	(0.295)
R^2	0.484	0.468	0.524
Observations	1034	1007	960

Robust standard errors in parentheses. ** $p < 0.01$, * $p < 0.05$.
Notes: Variables are the same as table 9.
Source: Same as table 9.

Figure 3: OBP and SLG Market Valuation: Hakes and Sauer (2006)

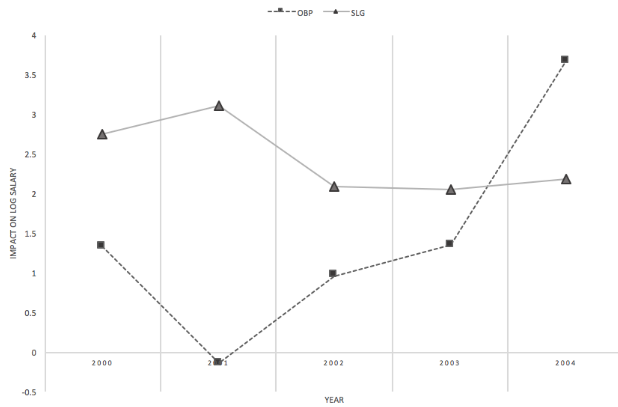
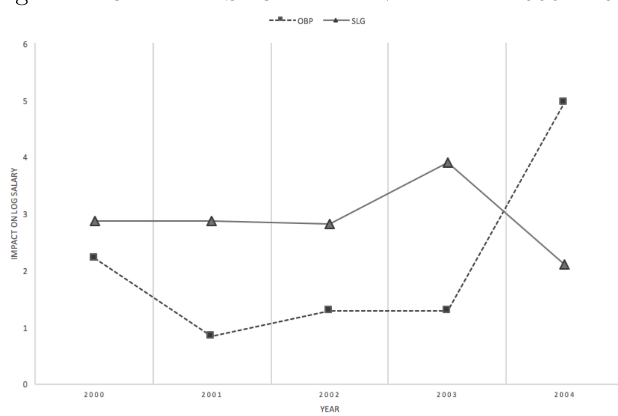


Figure 4: OBP and SLG Market Valuation: 2000 - 2004



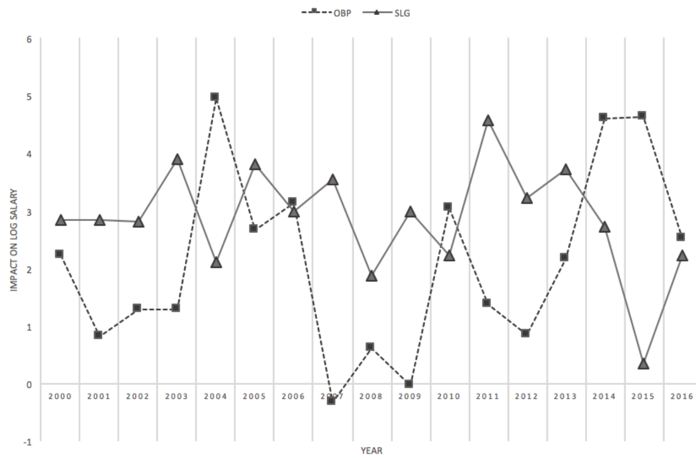
3.1 Can the labor market regressions of Hakes and Sauer (2006) be replicated?

The results of Hakes and Sauer's regressions is shown in table 8, and the replication results in table 9. The signs on all coefficients match that of Hakes and Sauer (2006). However, the differing amount of observations show that there is a clear discrepancy in the size of the sample populations. This margin of error is most likely caused by the lengthy data collection process that is inherent

[6] Arbitration takes place after a player has served between 3 and 6 full years in the MLB. At this time, a player is able to enter arbitration with the contract holding team in order to negotiate a more "fair" salary. This is overseen by a mediator, who, unless a player and team settle outside of arbitration, evaluates the player's performance and calculates a new salary.

[7] After 6 full years of service, a player is able to sign with whatever team they want, and is thus a "free agent".

Figure 5: OBP and SLG Market Valuation: 2000 - 2017



to the data sets gathered from the SeanLahmanDatabase (2019). The process of merging large data sets, and dropping duplicate observations caused by in-season trades and multi-positional players, most likely led to differing sample population sizes between the original and replication data sets. The differing in sample sizes also has an effect on the coefficients. In all columns, the coefficients on the key independent variables of OBP and SLG differ between the original and replication results by small margins. For the 2000 - 2004 seasons, the replication results differ from the results of Hakes and Sauer (2006) by 0.414 and 0.506 for OBP and SLG respectively. For 2000 - 2003, the difference is 0.292 and 0.636 for OBP and SLG respectively. And for the 2004 season, the difference between coefficients is 1.284 and 0.059 for OBP and SLG respectively. This differentiation is almost certainly due to the differing amounts of observations in each regression.

That said, it is still apparent that the replication is relatively successful in reproducing the results of Hakes and Sauer (2006). The first regression, seen in column one of both tables 8 and 9, shows the salary estimates for the years 2000 - 2004. In column two, using data from the four years leading up to the publishing of *Moneyball* in 2003, both tables show OBP decreasing and SLG increasing. However, for both of these columns, only the coefficient on SLG is significant. In 2004, the year after *Moneyball* was published, the replication is successful in corroborating the findings of Hakes and Sauer by displaying a large increase in the valuation of OBP, and a large decrease in the valuation of SLG. Moreover, in this column, both the OBP and SLG coefficients are statistically significant. From these results it can be observed that the replication successfully confirms the *Moneyball* hypothesis suggested in Hakes and Sauer (2006) by showing an undervaluation of OBP leading up to 2004, and a reversal of the valuation after 2004. Hakes and Sauer's findings of the OBP and SLG valuation trend between the years of 2000 - 2004 can be seen in figure 3. The coefficient on OBP surpasses that of SLG in the year 2004, with OBP jumping from 1.351 to 3.681.

The same trend is shown in the graph of the replication results, found in figure 4. Here, OBP increases from 1.298 to 4.965, far exceeding the value that teams put on SLG. This may be an example of a temporary overreaction by the market in its correction of the previous inefficiency. This is overshooting, and is discussed in further detail later in the paper. No matter the cause, the replication is successful in showing the key trend of Hakes and Sauer (2006)—the reversal of valuation for OBP and SLG in the MLB player market following the 2003 season.

The extension of Hakes and Sauer's study can be found in table 10, which uses data from the 2000 - 2017 seasons. The first column shows data from the 2005 - 2008 seasons. The coefficients on OBP and SLG point to their valuation returning back to levels that were found before the release of *Moneyball* in 2003. However, the coefficient on OBP is not significant. The 2009 - 2012 seasons show similar results, with the value of OBP decreasing by even more, and SLG rising slightly. Column three shows the most up-to-date salary data that is currently available, which spans to the 2017 season. This estimate shows OBP jumping back up to 3.566, and SLG dropping to 2.123—both of which are significant at the one percent level. The valuation trends on a yearly basis is shown in figure 5. It is clear that the valuation of both OBP and SLG is very volatile from year to year. The value of the two metrics cross each other multiple times, with no obvious trend or pattern.

Following the results of Hakes and Sauer (2006), it was expected that the MLB labor market had permanently corrected itself—assigning proper value to OBP as a more effective predictor of WPCT than SLG. However, as figure 5 clearly shows, this does not appear to be the case. The market value of OBP and SLG seems to completely differ from one year to the next. Following the publication of *Moneyball*, MLB front offices may have bought into the strategy of the Oakland Athletics, as was posited in Hakes and Sauer (2006). Or it is possible that the valuation of OBP in 2004 jumped to such a high level by mere chance. In any case, it is evident that this trend did not sustain even to the following year of 2005. Teams began to, once again, assign value to OBP and SLG with seemingly little regard to their actual contribution to wins.

However, this is not all that unexpected. The nature of a free-agent market during any given year naturally produces different results in terms of salary data. The following section discusses this in greater depth.

4 Internal and External Validity

4.1 Are the regressions of Hakes and Sauer (2006) internally valid?

One of the most common and detrimental threats to the internal validity of a regression is *reverse causality* or simultaneous equation bias (Stock and Watson, 2019). This occurs when the dependent variable causally affects one or more of the independent variables. The presence of this causes the independent variable in question to be correlated with the error term. Reverse causality produces inconsistent and biased results. When examining Hakes and Sauer's first regression, found in table 2, there does not appear to be a presence of reverse causality. It may seem reasonable that teams with high levels of WPCT are more likely to have high payrolls, and in turn be able to purchase players with high OBP and SLG. However, the lag effects of compensation in professional sports protects against this. Because players are usually evaluated on their prior performance, i.e., the previous year or years' statistics, and not their current production, there is no way for WPCT to affect a team's level of OBP and SLG. This is also the case with Hakes and Sauer's second regression, found in table 8. Log salary has no impact on a players OBP and SLG from the year prior, nor does it affect any of the other independent variables in the regression.

Specification error is also a threat to internal validity. Linear regression using OLS is adequate for the regression of WPCT on OBP and SLG. There is no evident reason suggesting that a nonlinear regression would be better suited for this circumstance, and the scatter plot of the regression seems to be well-matched for the currently utilized specification as can be seen in figures 1 and 2.

Hakes and Sauer's second regression uses a log-linear regression function. Using this specification makes sense in this instance because it transforms a player's salary into percentage terms. One million dollars paid to a player who plays for the Tampa Bay Rays[8] does not have the same meaning or significance as one million dollars paid to a player who plays for the Boston Red Sox.[9] Along with this, different positions earn different salaries based on their market value. A short-stop may be more valuable than a right fielder because the position is harder to play defensively. Even if he is a worse hitter, the short-stop may be paid more because of the value added to winning percentage. It is easier to compare hitters across differing teams and positions, through percentage increases rather than in dollar terms, which makes the log-linear specification the proper form for this regression.

Probit, logit, and linear probability model (LPM) regressions are run on the data set used in tables 9 and 10. This is done in order to test whether any of these models are better predictors of the effect of OBP and SLG on salaries in comparison to Hakes and Sauer's log-linear model. To construct these regressions, the dependent variable of log salary is changed to a binary variable which equals 1 for high log salaries (above the mean), and 0 for low log salaries (below the mean). The results of these regressions can be found in the appendix. The coefficients on OBP, one of two key variables in the regressions, are statistically insignificant for all three models; indicating that probit, logit, and LPM are not reliable estimators in this circumstance. This suggests that

[8]The Tampa Bay Rays have perennially had one of the lowest payrolls in the MLB. Their 2019 payroll was last in the league at 53.4 mn USD(BaseballProspectus, 2019).

[9]The Boston Red Sox's 2019 payroll was 202.9 mn USD, more than 280 percent higher than that of the Tampa Bay Rays (BaseballProspectus, 2019).

there is not a misspecification of functional form in Hakes and Sauer (2006).

Measurement error is another a threat to internal validity. The measurement of the independent and dependent variables found in Hakes and Sauer's first regression is straightforward. Team statistics are consistent and easy to collect. The process of assembling the data for the second regression, however, is more nuanced. As noted above, when large sets of data are merged, there are many duplicates on the player level due to in-season trades and multi-positional players. The ensuing process of identifying and purging duplicates is lengthy and error prone. As can be seen above, the replication results produced an inconsistent number of observations to Hakes and Sauer (2006). Also, this study was not able to reproduce the variables *Arbitration eligible* and *Free agency* with any accuracy. To combat this, the variable of *Age* was used in their place. The difficulty of determining whether a player is a free agent or eligible for arbitration points to the possibility of measurement error within the original data set of Hakes and Sauer (2006). Given the database that was used by Hakes and Sauer, the methodology for determining free agency and arbitration eligibility would have to derive from predicting a player's service time based on their debut date. This can be inaccurate as it is common for MLB teams to manipulate the service time of young players in an attempt to stave off their eligibility for arbitration. While Hakes and Sauer unquestionably implemented a careful and efficient approach to determining these variables, an effort to reproduce their data for replication uncovered a number possible sources of measurement error. For ease of replication, it might be more beneficial for the authors to have used the variables *age* and age^2 as measures of a player's bargaining position. Even though *Arbitration eligible* and *Free agency* are not key variables, this possible measurement error is a threat to the internal validity of the regression.

Omitted variable bias is yet another a threat to internal validity. Here are no apparent signs of omitted variable bias in either regression. In the OLS linear regression estimating the effect of OBP and SLG on WPCT, it is important to add a variable which accounts for a team's pitching competence, and the opposing team's hitting proficiency. This is important because pitching makes up a large portion of why a team is a winning or losing team. It could be contended that fielding percentage, along with hitting and pitching, is also important to a team's success. However, most baseball analysts and sabermetricians have little confidence in defensive metrics. They are often highly subjective and provide limited insight into how proficient a defender is (Basco and Zimmerman, 2010). In their second regression determining the causality of various variables on log salary, Hakes and Sauer account for the hitting ability, service time, and position of each player.[10] This seems to be sufficient for the model.

Sample selection bias in observational studies is rare, especially for experienced researchers. In selecting their sample, Hakes and Sauer used all MLB non-pitchers with over 130 at-bats. This 130 at-bat threshold is a reasonable amount for a player to produce statistics that are indicative of their ability. For this reason there does not seem to be any signs of selection bias in Hakes and

[10]Omitting outfielders to avoid the dummy variable trap.

Sauer (2006).

The models of Hakes and Sauer (2006) appear to be internally valid. The only threat to their internal validity is the possibility of measurement error within the second regression. However, there is a lack of substantive evidence to suggest that this measurement error is egregious, or that it biases the results in one direction or the other in a major way.

4.2 Are the regressions of Hakes and Sauer (2006) externally valid?

Although the results of Hakes and Sauer (2006) are consistent and valid when pertaining to the 1999 - 2004 sample, they are not externally valid when the sample population is expanded. The model presented in Hakes and Sauer (2006), as well as the replication results, show that prior to the 2004 season SLG was consistently awarded more value than OBP by MLB teams; with this trend reversing after the 2004 season. However, this generalization cannot be expanded beyond the 2004 season. As can be seen in figure 5, the valuation of OBP and SLG fluctuates, and is extremely volatile over the course of the 2000 - 2017 seasons. No distinct trend in how teams award value to the two metrics is apparent. This points to the statistical inferences offered in Hakes and Sauer (2006) as being externally invalid. Although there is a valuation reversal in 2004, it is one of many such reversals in the MLB labor market. After zooming out, and examining a wider range of data, it is difficult to say whether the release of *Moneyball* in 2003 had any effect on how MLB front offices value players or not.

There are many possible reasons for the presence of such volatility in the MLB player market. The fluctuation in salaries on a year-by-year basis is largely driven by the free-agent market. Intuitively, overall player compensation in the MLB rises and falls because of the singing of new contracts. Most of these new signings are free agents whose contracts have expired and are either looking for a new team, or resigning with their original team. The free-agent market for any given year will be made up of a random mix of players, with some years possessing a high volume of power-hitting free agents, some years possessing a high volume of hitters with high on-base percentage, and so on. If a free-agent market has an abundant amount of top-tier power-hitters, then the average compensation level for players with high SLG is going to increase. If the next year's free agent market is filled with high-level hitters who get on-base efficiently, then the average salary for a player with high OBP will increase. Thus the MLB valuation of hitters can easily change from one year to the next simply depending on the random make-up of the free agent market. This would explain the amount of fluctuation in the value of OBP and SLG that can be observed in figure 5.

Another possible explanation, which was briefly mentioned earlier in the paper, is the economic concept of overshooting. It could be the case that, after the release of *Moneyball*, MLB teams *did* realize that there was an inefficiency in the player market, and over-corrected their valuations of OBP and SLG in an attempt to fix it. Following this, as teams tried to return to an equilibrium state of properly valuing the two metrics, ensued a period of extended instability in the market. This is common in money markets where currencies are subject to changes in valuation due to floating exchange rates (Bjørnland, 2009).

Either way, it is clear that the findings of Hakes and Sauer (2006) are externally invalid. Although the statistical inferences that can be made from their regressions are pertinent to the original sample population, their causal effects are no long valid when the population is expanded.

5 Conclusion

This paper replicates the methodology and findings of Hakes and Sauer (2006), as well as cross-validates their results with an agent-based model. The regressions of Hakes and Sauer (2006) show that OBP affects WPCT more than SLG does. The replication and cross-validation of these findings were successful in corroborating this. Their study proceeds to suggest that leading up to the 2004 season, OBP was undervalued in the MLB labor market, which was then corrected soon after the publication of *Moneyball* the previous year. Due to different data compiling methodologies, this paper is ineffective in producing the exact results of Hakes and Sauer (2006); however, it is successful in replicating the trends found in Hakes and Sauer (2006) and confirming their findings. Most coefficients are very close to that of Hakes and Sauer (2006), and the market valuation trends of OBP and SLG prove extremely similar.

This paper finds the models of Hakes and Sauer (2006) to largely be internally valid. However, upon extending Hakes and Sauer's study, the results suggest that the generalizations made in their article cannot the expanded to a larger sample population. For that reason, this paper determines their estimates to be externally invalid, and suggests a new outlook on the *Moneyball* hypothesis. While the relative value of OBP to SLG was corrected after the 2003 season, the valuation did not remain correct for long. Since then, the market value of OBP and SLG has varied greatly—with no semblance of equilibrium being reached. More than anything, the valuation of OBP and SLG during any given year is more likely to be a result of the makeup of the prior year's free agent market.

6 Pseudo Code

The program written for the agent-based model can be expressed as:

Set Parameters

1. Create 10 teams (patches)
   (a) Set team on-base percentage random 0.292 to 0.373
   (b) Set the probability of a single 0.71

(c) Set the probability of a double 0.16

(d) Set the probability of a triple 0.009

(e) Set the probability of a home run 0.12

2. Create 20 players (turtles) and assign to random team

    (a) Set players batting statistics equal to team statistics

3. Set total runs to 0

4. Set total Wins to 0

5. Set total losses to 0

6. Set total ties to 0

Match-Up Sequence

1. Ask random team to play a different random team who has played < 162 games

2. Set one team home and one team visitor

Play Game

1. Play half inning with visitor hitting

    (a) Play at-bat sequence

        i. Set pitch as random decimal 0 to 1

        ii. If pitch $>$ OBP batter is out

            A. If less than two outs and runner on first set double play probability 0.15

        iii. If pitch $<$ OBP batter gets a hit

            A. If hit is a single: player goes to first, runners advance one base, runners on third score, probability runners on second score $= 0.4$

            B. If hit is a double: player goes to second, runners advance two bases, runners on second and third score, probability runners on first score $= 0.14$

            C. If hit is a triple: player goes to third, all runners score

            D. If hit is a home run: team scores on run + however many runners are on base, bases clear

    (b) Total the runs from this inning and add to score of visiting team

2. Play half inning with home team hitting

    (a) Repeat at-bat sequence

    (b) Total the runs from this inning and add to score of home team

3. Repeat for nine innings
4. If score of home team > or = score of away team at the end of nine innings
    (a) Give corresponding teams one loss, win, or tie

Play Season

1. Repeat match-up sequence until all teams have played 162 games

Record Seasons statistics for each team

1. Record wins
2. Record losses
3. Record ties
4. Record batting statistics

Repeat for a total of 21 seasons

7 Appendix

Table 11: Parameter Values for the Agent-Based Model

Team Endowment Parameters		
r	Runs	0
w	Wins	0
l	Losses	0
t	Ties	0
1_b	Single Percentage	0.71
2_b	Double Percentage	0.16
3_b	Triple Percentage	0.009
H_r	Home run Percentage	0.12
O_{bp}	On-Base Percentage	0.292 - 0.373
O_{bp}	On-Base Percentage	0.335 - 0.495

Table 12: Probit, Logit, and LPM: Salary Data 2000 - 2017

Salary Dummy[1]	Probit (1)	Logit (2)	LPM (3)
OBP	1.041	1.699	0.083
	(0.805)	(1.374)	(0.215)
SLG	3.286***	5.639***	0.876***
	(0.436)	(0.754)	(0.116)
Plate appearances	0.004***	0.008***	0.001***
	(0.000)	(0.000)	(0.000)
Age	0.142***	0.252***	0.040***
	(0.007)	(0.013)	(0.002)
Catcher dummy	0.140*	0.250*	0.043*
	(0.069)	(0.118)	(0.020)
Infielder dummy	-0.091	-0.161	-0.024
	(0.049)	(0.085)	(0.013)
Constant	-7.941***	-13.834***	-1.689***
	(0.305)	(0.560)	(0.061)
Pseudo R^2	0.306	0.308	
R^2			0.354
Observations	4379	4379	4379

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Notes: 1. Binary log salary variable that equals 1 when log salary is above the mean, 0 when below the mean. The rest of the variables are the same as table 9.
Source: Same as table 9.

References

Basco, D. and J. Zimmerman (2010). Measuring Defense: Enter the Zones of Fielding Statistics. *Society for American Baseball Research*.

Baseball-Reference (2019). 2019 MLB Team-Batting Statistics. *https://www.baseball-reference.com/*.

BaseballProspectus (2019). Cot's Baseball Contracts.

Bjørnland, H. C. (2009, September). Monetary Policy and Exchange Rate Overshooting: Dornbusch was Right After All. *Journal of International Economics 79*(1), 64–77.

Hakes, J. K. and R. D. Sauer (2006). An Economic Evaluation of the Moneyball Hypothesis. *The Journal of Economic Perspectives 20*(3), 173–186.

Issadore, K. (2009). The Moneyball Hypothesis in Major League Baseball. The University of Vermont.

Lewis, M. (2003). *Moneyball: The Art of Winning an Unfair Game.* W. W. Norton and Company.

MLB.com (2019). On-base Plus Slugging Percentage Plus (OPS+). *http://m.mlb.com/glossary/advanced-stats/on-base-plus-slugging-plus*.

SeanLahmanDatabase (2019). 2019 Fielding, Salary, Batting, and Master Data Sets. *http://www.seanlahman.com/baseball-archive/statistics/*.

Stock, J. H. and M. W. Watson (2019). *Introduction to Econometrics* (4 ed.). The Pearson Series in Economics. Pearson.