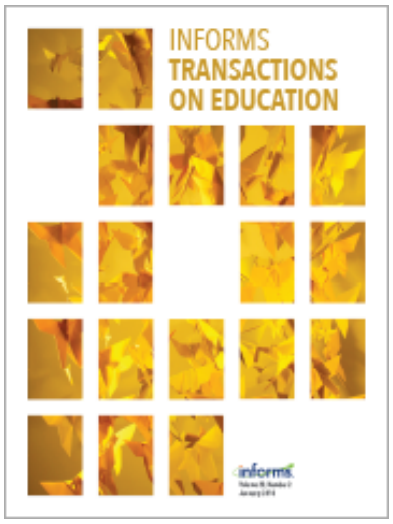


INFORMS Transactions on Education

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Case—Baseball Analytics: Advancing to Prescriptive Analytics in the Major League Baseball Front Office

Sean L. Barnes, Margrét V. Bjarnadóttir

To cite this article:
Sean L. Barnes, Margrét V. Bjarnadóttir (2019) Case—Baseball Analytics: Advancing to Prescriptive Analytics in the Major League Baseball Front Office. INFORMS Transactions on Education 19(3):146-151. https://doi.org/10.1287/ited.2018.0201cs

http://pubsonline.informs.org/journal/ited/

INFORMS TRANSACTIONS ON EDUCATION
Vol. 19, No. 3, May 2019, pp. 146–151
ISSN 1532-0545 (online)

Case—Baseball Analytics: Advancing to Prescriptive Analytics in the Major League Baseball Front Office

Sean L. Barnes,[a] Margrét V. Bjarnadóttir[a]

[a] Department of Decision, Operations and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742-1815

Contact: sbarnes@rhsmith.umd.edu, http://orcid.org/0000-0001-5497-6277 (SLB); margret@rhsmith.umd.edu, http://orcid.org/0000-0003-2955-1992 (MVB)

Received: October 27, 2017
Accepted: May 28, 2018
Published Online in Articles in Advance: May 20, 2019

https://doi.org/10.1287/ited.2018.0201cs

> Intelligence about baseball statistics had become equated in the public mind with the ability to recite arcane baseball stats. What [Bill] James's wider audience had failed to understand was that the statistics were beside the point. The point was understanding; the point was to make life on earth just a bit more intelligible; and that point, somehow, had been lost. "I wonder," James wrote, "if we haven't become so numbed by all these numbers that we are no longer capable of truly assimilating any knowledge which might result from them."
> (Lewis 2003, p. 95)

Each Major League Baseball (MLB) team has a general manager (GM) who primarily oversees personnel decisions and negotiations concerning the acquisition, assignment, and release of players and coaches. MLB GMs and their staffs are charged with identifying and acquiring players for the team via (a) trades with other teams, (b) the free agent market, through which teams compete to sign eligible players whose previous contracts have expired, (c) the team's minor league development system,[1] or (d) an annual amateur draft, during which players from high schools, junior colleges, universities, and international markets are selected and signed to play in either a team's minor league system or at the major league level. In recent years, professional players outside of the United States have also been signed to major and minor league contracts.

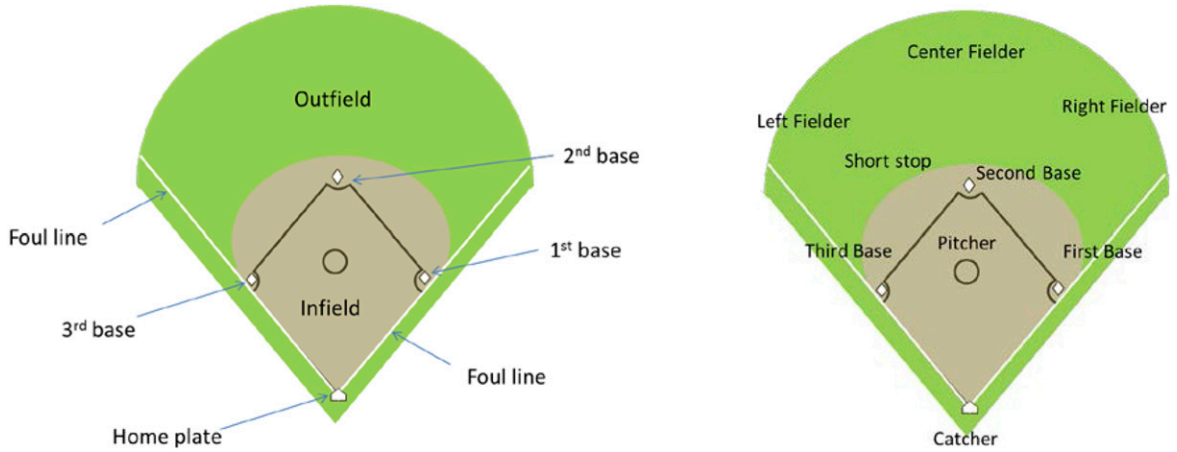
When attempting to sign a player via free agency, the GM must determine the value of that player—in terms of their contributions on and off the field—and then negotiate a contract with that player's agent. How does on-the-field performance translate to a player's contract value (i.e., salary)? That is a question that baseball executives need to answer when determining how to field a competitive roster for the upcoming year. These types of signing decisions are usually subject to financial constraints.

Baseball Basics[2]

Professional baseball, as we know it today, originated in the mid-1800s. The game is played between two teams who compete to score the most runs over the course of nine innings. Each inning is divided into two halves, a top half and a bottom half. In the top half of an inning, the visiting team is on offense, whereas the home team is on defense. During a half inning, the team on offense attempts to score as many *runs* as possible by advancing runners around the bases. The team on defense attempts to prevent runs from being scored by recording three *outs*, which ends the half inning. A defensive team records an out when they retire (i.e., remove) an offensive player from the bases (or prevent one from reaching a base). There are many ways that teams on offense can advance players around the bases and many ways that teams on defense can get players out. At the end of each half inning, the teams switch sides and play through at least eight-and-a-half innings[3] or through additional innings until a winner has been decided.[4] Major League Baseball in the United States is organized into two leagues, the National League (NL) and the American League (AL). In the NL, the same nine players play offense and defense, whereas in the AL, a designated hitter bats in place of the pitcher.

The baseball field is composed of an infield and an outfield (see Figure 1). In the infield, there are four bases—first base, second base, third base, and home plate—in a diamond shape and a mound in the middle, where the pitcher pitches the ball toward a batter standing at home plate. Six players play defense on

Figure 1. Diagram of a Baseball Field



Notes. (Left) Key parts of the field are labeled. (Right) Defensive positions and their approximate locations on the field are labeled.

the infield (known as infielders, except the pitcher and catcher), and three players play defense in the outfield (outfielders). Foul lines extend from home plate through first base and third base to divide fair territory—where all batted balls are in play—from foul territory. Outs can be recorded in foul territory when a defensive player catches a batted ball before it hits the ground (i.e., a foul out).

On offense, players have one of three roles: (a) batting, (b) base running, or (c) waiting to bat. During an *at bat*, the batter attempts to get on base, either via a *hit* or a *base on balls* (also referred to as a *walk*). A batter gets a hit when he hits the ball into fair territory and safely reaches base. *Singles* are hits with which the player reaches first base. *Doubles*, *triples*, and *home runs* are hits with which the batter reaches second base, third base, and home plate, respectively, on the same play. Each batter is allowed multiple attempts to bat the ball into fair territory but is limited by the number of strikes, which are pitches that are (a) swung at and missed by the batter, (b) thrown into a standardized strike zone, or (c) batted into foul territory (except when there are already two strikes). If the batter takes (i.e., does not swing at) four pitches outside of the strike zone—called *balls*—he earns a walk and gets to advance to first base. Batters can also reach base if they are *hit by a pitch* or if a defensive player commits an *error*.

If a batter does reach base, he then becomes a base runner during the subsequent at bats by his teammates during the half inning. There is only one base runner allowed per base, therefore base runners need to vacate their current base if a subsequent batter or base runner attempts to run toward it. Base runners do not have to run if there is a vacant base behind them. Base runners can potentially advance around the bases in several ways, including on batted *balls in play* (i.e., hits or outs), walks, hit by pitches, or by stealing a base or advancing on a pitch that is not caught by the catcher (e.g., a wild pitch or passed ball). Each base runner that reaches home plate during the half inning scores a run for his team.

On defense, a team's pitcher and the other fielders attempt to get any combination of three batters and base runners out while minimizing the number of runs scored by the offense. The most common ways to get a batter out are via strikeouts, ground outs, and fly outs. A pitcher strikes out a batter when three strikes are thrown in any combination during an at bat. A batter *grounds out* when he hits the ball on the ground and the ball is thrown to first base before the batter arrives. A batter *flies out* when a fielder catches his batted ball before it lands on the ground. The defense can also get base runners out in several ways, most commonly by *tagging out* (i.e., by touching the base runner with the ball) or *forcing out* (i.e., touching a base to where a batter is forced to run) base runners trying to reach a particular base. Additional references for the rules of baseball can be found in Appendix A.

Baseball Statistics

Baseball statistics have been collected since the beginning of organized play. The most basic statistics are simple counts of significant events for each player. For batters, the most basic statistics include information on the number of times the player batted and the results of those at-bats. The number of times a batter bats is called plate appearances (PA), whereas the official at-bat statistic (AB) excludes unofficial at-bats, such as when the batter walks, is hit by a pitch, or sacrifices himself to advance the base runners (e.g., a sacrifice bunt [SH] or sacrifice fly [SF]). The other basic batting statistics are the number of hits (H)—including the

specific number of doubles (2B), triples (3B), and home runs (HR)—and the respective number of walks (BB), hit by pitches (HBP), and strikeouts (K). There are also statistics for base runners, pitchers, and defensive players. A list of additional references for statistics can be found in Appendix A, and a list of the most common statistics and their abbreviations is provided in Appendix B.

In addition to simple count statistics, there are also statistics that attempt to capture the efficiency of players (i.e., how likely they are to generate a specific outcome in a single attempt). These statistics facilitate comparison between batters with a different number of at-bats or pitchers with a different number of innings pitched. For batters, the most common statistic of this type is batting average (BA or AVG), which is the number of hits per at-bat (H/AB). By traditional standards, a good batter has a 0.300 (pronounced "three hundred") average, which means that he averages three hits for every ten at bats. Another statistic for batters is slugging percentage (SLG), which attempts to capture the impact of a batter's hits, in terms of the average number of bases the batter advances per hit. This aspect is ignored in the batting average, which only captures the proportion of at bats resulting in any type of hit. SLG is calculated as follows:

$$SLG = \frac{(1B) + 2(2B) + 3(3B) + 4(HR)}{AB} = \frac{TB}{AB}.$$

The numerator of SLG is the number of total bases (TB), a weighted sum that rewards one base for each single, two bases for each double, three bases for each triple, and four bases for each home run. Although BA, SLG, and TB capture batting efficiency better than simple counts, they still have limitations in capturing a batter's entire contribution to his team's ability to score runs. For example, neither statistic captures walks or sacrifices, which can also help a team score runs. A statistic that takes walks and hit by pitches into account is called on-base percentage (OBP), which is the fraction of plate appearances for which a player reaches base. OBP is calculated as follows:

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF}.$$

The members of the Society for American Baseball Research (SABR), as well as academic researchers, practitioners, and fans of the sport have spent tremendous efforts on researching baseball statistics, including creating new statistics and analyzing how statistics both new and old translate into performance on the field and the probability of winning games and championships.

In the 1990s, the SABR influenced the adoption of the *slash line*, which lists a player's AVG, OBP, and SLG separated by slashes. For example, a player with an excellent slash line would have an AVG of 0.300, an OBP of 0.400, and an SLG of 0.500, which would be represented in slash line form as 0.300/0.400/0.500. The slash line provides a more robust representation of a player's performance than any single statistic. As useful as the slash line can be, the general preference remains to evaluate the overall value of a player in a single statistic. Multiple statistics have been proposed. One of the most widespread SABR metrics—popularly referenced as sabermetrics—is OBP plus SLG, or OPS. OPS attempts to capture the ability of a player to get on base and the ability to hit for extra bases. Another example is wins above replacement (WAR), which is a statistic that attempts to capture the overall value of a player relative to a replacement player who is below average offensively and average defensively. This statistic has quickly become one of the most important measures of overall player performance. We direct the interested reader to Tymkovich (2012) for an overview of WAR and other overall performance metrics.

Data Analytics in Baseball

Although statistics have a long history in baseball, using advanced statistics to guide hiring decisions is a relatively recent phenomenon. The inspiring performance of the Oakland Athletics (A's) during the 2002 season produced an increased focus on analytics in baseball and inspired the immensely popular book *Moneyball: The Art of Winning an Unfair Game* (Lewis 2003), which in 2011 was released as a film featuring actor Brad Pitt. The primary strategy of the A's was to identify undervalued players to build a successful yet inexpensive team. In other words, they wanted to identify statistics that were relevant to success on the field but overlooked when players were evaluated for a contract. At the heart of Oakland's success was an observation that players with higher batting averages and slugging percentages (i.e., traditional statistics) were overvalued, whereas players with the ability to get on base frequently were undervalued in the player market. By acquiring players with these undervalued attributes, the team won more games than what would be predicted by its payroll.

You are a new hire to the management staff for a small-market MLB team. You have been tasked with understanding changes in player evaluations as a result of this new wave of data analytics. In particular, a summer intern has collected statistics and salary information for all MLB free agent batters from the 1998 season through the 2013 season. You plan to use these data to accomplish your task. Your goal is to submit a report to the head of the analytics department that the team can use in their free agent negotiations. This report should describe how players are currently evaluated, how player evaluation has changed as a result of more widespread use of data analytics, and what can be considered a "fair" salary for free agents.

Part A

You realize that to devise a salary prediction model, you first need to understand the data. Take some time to become familiar with the data; specifically, you should understand how the data are structured and gain a basic understanding of the types of statistics contained in the data. After gaining a thorough understanding of the data, you plan to build a model that best represents the salaries of free agent batters. You wonder whether you should build the best prediction model or the best explanatory model, or perhaps there is no difference?

Part B

You recognize that you are not the first person to take on the challenge of analyzing the changes that *Moneyball* brought to the sport of baseball. In particular, you learn of two published studies that have attempted to quantify the effect of the *Moneyball* methodology on player salaries: a paper by authors Hakes and Sauer (2006) and a web article published by Farrar and Bruggink (2011).

In their study, Hakes and Sauer confirm the market inefficiencies suggested by the *Moneyball* hypothesis. In particular, they show that player salaries (excluding pitchers) were not reflective of their contributions to wins. First, the authors confirm with linear regression analysis that a high OBP has a strong positive effect on winning games—stronger than the effect of SLG—and, therefore, a player's salary should have a stronger link to OBP than SLG. To show how the evaluation of these two statistics changed after the success of the Oakland A's in 2002, they use data on all nonpitching players with more than 130 ABs in a season. They use the logarithm of next year's annual salary as the dependent variable and build a model to explain it with seven independent variables in their model: OBP, SLG, and PA, as well as binary variables indicating whether the player was eligible for salary arbitration (i.e., between three and six years of experience), a free agent (i.e., more than six years experience), a catcher, or an infielder. They first run a regression on the data across all years and find that the SLG coefficient is larger than the coefficient for OBP. When the analysis is run on individual years, the SLG coefficient is larger than the OBP coefficient in the pre-*Moneyball* years (2000–2003), but the OBP coefficient is larger in the last year (2004). They conclude that their result "is consistent with *Moneyball*'s claim that on-base percentage is undervalued in the labor market"; however, they acknowledge some limitations of the study, such as ignoring long-term contract effects and the fact that winning may not be a team's only objective (e.g., it could be overall profit maximization).

Farrar and Bruggink (2011) argue that Hakes and Sauer overestimate the diffusion of the *Moneyball* analytics. They claim that using all players instead of only recently signed players limits the analysis because it dilutes the player evaluation. Player contracts are typically multiyear contracts; therefore, even if the effect of *Moneyball* was instantaneous it would not be fully reflected in salaries until several years later. To overcome this limitation, the authors only include experienced free agent players that play regularly in their data set. The authors use two regression models. They first model the contribution of OBP and SLG to runs, because they argue that batters are paid for producing runs, not necessarily wins. If players are "fairly" evaluated, then they would get paid in the same proportion as these performance metrics contribute to runs. Fitting their second model, the authors find that SLG is still overvalued three years after the release of *Moneyball* when compared with OBP.

Using your data set, devise your own model for evaluating the diffusion of *Moneyball* analytics into MLB. How do your results compare with the results of these two articles?

Part C

Now that you have analyzed some key research in this area, incorporate the full extent of your data set into your model for player salary. Three things you should consider when building your model:

- What characteristics of your model are most important when evaluating its quality?
- Variable selection: What variables should be included? Are there certain variables that should not be included together? How does the number of variables affect the quality and interpretability of your model?
- Data segmentation: Should you run the model on the entire data set, or should you subset the data and fit separate models for each subset?

After running your model, take some time to translate your results into valuable insight for the director of analytics for your team. Specifically, you should address:

- Which player statistics have the most impact on salaries?
- Which players were the most under- and overvalued according to your model?

Part D

Finally, now that you have gained an understanding for how players have been compensated for their on-the-field performance, devise a model for how players should be compensated given the same considerations (i.e., the same available data). How would you use your models to support free agent selection?

Appendix A. Additional Resources

The Flow of the Game

- Understanding the game through cartoons: http://video.disney.com/watch/how-to-play-baseball-4be388191e056a0e1266b068

Baseball Basics

- How Stuff Works: http://www.howstuffworks.com/baseball.htm
- How Baseball Works: http://www.howbaseballworks.com/
- Wikipedia: http://en.wikipedia.org/wiki/Baseball
- A Public Broadcasting Service guide: http://www.pbs.org/kenburns/baseball/beginners/

Basic and Advanced Baseball Statistics

- Baseball Stats 101: http://www.baseball-almanac.com/bstatmen.shtml
- Advanced Stats: http://www.fangraphs.com/library/
- Baseball Reference: http://www.baseball-reference.com/; navigate to any player or team page and utilize the tool tips by hovering over any statistic or clicking the Glossary, which describes the stats in each table.
- Baseball Guru: http://baseballguru.com/bbguruol.html

Evaluating the *Moneyball* Effect

- Farrar A, Bruggink TH (2011) A new test of the *Moneyball* hypothesis. *The Sport Journal*, 14. Available at: http://thesportjournal.org/article/a-new-test-of-the-moneyball-hypothesis/
- Hakes JK, Sauer RD (2006) An economic evaluation of the *Moneyball* hypothesis. *J. Econom. Perspect.* 20(3):173–185
- Sabermetric research, for example:
  - http://blog.philbirnbaum.com/2006/10/how-fast-did-market-learn-from.html
  - http://blog.philbirnbaum.com/2007/10/updated-moneyball-effect-study.html

Appendix B. Commonly Used Baseball Statistics and Abbreviations

Offensive statistics		Pitching statistics		Defensive statistics	
AB	At bats	BB	Bases on balls (walks)	A	Assists
BB	Bases on balls (walks)	BF	Batters faced	dWAR	Defensive wins above replacement
AVG	Batting average (BA)	BK	Balks	DP	Double plays
CS	Caught stealing	CG	Complete games	E	Errors
2B	Doubles	ER	Earned runs allowed	PB	Passed balls
GIDP	Grounded into double play	ERA	Earned run average	PK	Pickoffs
HBP	Hit by pitch	GF	Games finished	PO	Putouts
H	Hits	GO/AO	Ground outs per fly outs ratio	TC	Total chances
HR	Home runs	GP	Games played	UZR	Ultimate zone rating
IBB	Intentional base on balls (walks)	GS	Games started		
ISO	Isolated power	H	Hits allowed		
LOB	Men left on base	HBP	Hit batters		
oWAR	Offensive wins above replacement	HR	Home runs allowed		
OBP	On-base percentage	IBB	Intentional bases on balls (walks)		
OPS	On-base plus slugging percentage	IP	Innings pitched		
R	Runs scored	IRA	Inherited runs allowed		
RC	Runs created	IPS	Innings per start		
RBI	Runs batted in	L	Losses		
SF	Sacrifice flies	OBA	Opponents' batting average		
SH	Sacrifice hits (bunts)	R	Runs allowed		
SLG	Slugging percentage	SHO	Shutouts		
SB	Stolen bases	SO (K)	Strikeouts		
SO	Strikeouts	SO/BB	Strikeout-to-walk ratio		
TB	Total bases	SV	Saves		
3B	Triples	W	Wins		
WPA	Win probability added	WAR	Wins above replacement		
		WHIP	Walks plus hits per innings pitched		
		WP	Wild pitches		

Sources. http://www.baseball-almanac.com/stats4.shtml, http://en.wikipedia.org/wiki/Baseball_statistics.

Endnotes

[1] The minor league system is a hierarchical organization of leagues in which players have the opportunity to develop their skills in the hopes of advancing to the major leagues. Most MLB players play through multiple levels of the minor league system before earning an opportunity at the major league level.

[2] For links to introductory baseball tutorials, please refer to Appendix A.

[3] There is no need to play the bottom of the ninth inning if the home team is leading after the top of the ninth inning.

[4] The longest extra-inning MLB game was decided in 25 innings.

References

Farrar A, Bruggink TH (2011) A new test of the *Moneyball* hypothesis. *Sport J.* Accessed December 15, 2018, http://thesportjournal.org/article/a-new-test-of-the-moneyball-hypothesis/.

Hakes JK, Sauer RD (2006) An economic evaluation of the Moneyball hypothesis. *J. Econom. Perspect.* 20(3):173–185.

Lewis M (2003) *Moneyball: The Art of Winning an Unfair Game* (Norton, New York).

Tymkovich JL (2012) A study of Minor League Baseball prospects and their expected future value. CMC Senior Thesis, Paper 442. Accessed December 15, 2018, http://scholarship.claremont.edu/cmc_theses/442.