5-21-2019

The Impact of Analytics in Professional Baseball: How Long Before Performance Improves

Lee Freeman
College of Business, University of Michigan-Dearborn, lefreema@umich.edu

The Impact of Analytics in Professional Baseball: How Long Before Performance Improves

Lee Freeman
University of Michigan–Dearborn
lefreema@umich.edu

ABSTRACT

This study extends previous analyses of winning percentage since 2014 for the 30 MLB teams and whether these performance data could have been predicted by the teams' analytics adoption. The study also includes stadium attendance as a secondary indicator of performance. Based on MLB teams' 2014 analytics adoption as reported by ESPN, there are statistically significant differences in teams' winning percentage, attendance percentage, and cumulative (multiple seasons combined) winning percentage when looking at performance data in the three subsequent years (2015-2017). The differences in winning percentage remain significant for multiple seasons, though with decreasing statistical strength. MLB teams should be aware that an immediate benefit from analytics adoption may not occur, but subsequent years may see stronger results. This aspect of analytics adoption is a critical aspect to analytics usage, and this potential lag effect should be considered when adopting new methods and techniques in any part of the organization.

Keywords

Analytics, baseball, analytics adoption, team performance, results lag.

INTRODUCTION

The primary driver behind the adoption of business analytics within organizations (or industries) is to improve organizational performance through one or more factors. These could be higher revenue, lower costs, better product placement, higher customer satisfaction, better strategic decision-making, etc. (Seddon, Constantinidis and Dod, 2012; Trieu, 2017). As with nearly every business investment in technology, there needs to be a return on investment in observable value in terms of efficiency, effectiveness, or performance. Unfortunately, a return on investment for business analytics is difficult to measure (McCann, 2014). And if the return on investment cannot be measured, it is difficult to justify the expense and resources. However, organizations continue to adopt business analytics across nearly all industries, including professional sports. Analytics has the potential to improve performance on the field as well as with player development, personnel decisions, practice/training methods, marketing, and ticket pricing (Maxcy and Drayer, 2014).

Professional sports teams, and baseball teams in particular, are using analytics in a variety of ways. Journals and conferences provide new techniques, methods, and measurements of baseball, many of which were initially created by individuals independent of the league or any specific team. These new measurements provide stakeholders (teams, managers, scouts, players, and fans) the opportunity to analyze and discuss patterns and trends based on descriptive data and to generate predictive models based on these data (e.g., Baumer and Zimbalist, 2015). Yet, the literature, models, statistics, and discussions fail to consider whether analytics usage impacts performance and provides a return on the investment. Winning percentage is arguably the most important on-field performance variable as winning games is the primary goal of any team and the most recognizable measure of success. Attendance percentage is an indirect but easily obtained measure of off-field performance, as attendance impacts revenues from ticket sales, concession sales, and merchandise/souvenir sales (Freeman, 2016), and attendance impacts on-field performance (Smith and Groetzinger, 2010). However, Freeman (2016) found no significant differences with analytics adoption impacting winning percentage or attendance for the 2014 seasons across the four, major U.S. sports leagues – MLB, NBA, NFL, and NHL.

Yet, baseball teams continue to adopt and implement analytics in many aspects of their operations (Baumer and Zimbalist, 2015; Lampe, 2015; Lindbergh and Arthur, 2016; Eustis, 2018). In the discussion, Freeman called for future research to "look at performance measures in future seasons (2015, 2016, and beyond), and assess the impact of the 2014 categorizations on future performance." More specifically, Freeman asked whether "significant differences in on-field and off-field performance arise in future seasons based on current analytics adoption levels [and whether] a measurable lag between adoption and performance" exists, leading to these research questions:

- Can performance improvements be observed after analytics adoption for professional baseball teams?
- How long before such improvements are observed?

BASEBALL ANALYTICS

At the heart of professional baseball (and sports in general) is the desire to win and to do so consistently. Tools and techniques, whether in recruiting, training, or game-play, that provide owners, managers, trainers, scouts, and players with an understanding of past performance and/or a predictive look at future performance are likely to receive attention (Alamar, 2013). Given the abundance of available data, it is not surprising that professional baseball has turned to analytics in the hope of making better decisions. Bill James is often credited with starting the analytics revolution in baseball in the late 1970s which has, over time, expanded to other professional sports. Slowly at first, but with greater intensity of late, analytics staff have increased (Lindbergh and Arthur, 2016). The league, teams, and other organizations are spending more time and money developing new metrics and gathering, analyzing, and interpreting the vast amounts of data (Baumer and Zimbalist, 2015; Eustis, 2018).

Measuring Adoption

While many teams are increasing their analytics commitment as measured by usage, staff size, or public statements regarding buy-in, much about the nature of analytics adoption and use remains secretive and proprietary. Still, there have been recent attempts at quantifying the analytics usage by professional baseball teams. Maxcy and Drayer (2014) assessed the overall adoption percentage of Major League Baseball at 97%. Based on team data, expert opinions, and evaluative data, ESPN (2015) released a comprehensive evaluation of all 122 teams across the four major U.S. professional sports leagues and categorized each team into one of five categories: 1-All-In, 2-Believers, 3-One Foot In, 4-Skeptics, and 5-Nonbelievers. These categorizations (see Table 1) were based on “the strength of each franchise’s analytics staff, its buy-in from execs and coaches, its investment in biometric data and how much its approach is predicated on analytics” (ESPN, 2015).

ESPN Category 1	ESPN Category 2	ESPN Category 3	ESPN Category 4	ESPN Category 5
Red Sox	Orioles	White Sox	Diamondbacks	Marlins
Cubs	Royals	Angels	Braves	Phillies
Indians	Dodgers	Brewers	Reds	
Astros	Mets	Giants	Rockies	
Yankees	Padres	Mariners	Tigers	
A’s	Blue Jays	Rangers	Twins	
Pirates	Nationals			
Cardinals				
Rays				

Table 1. Analytics Categorizations of Professional Baseball Teams (ESPN, 2015)

Ferrari-King (2016) listed the top analytics teams across the four major professional sports and the honorable mention teams. The top teams (eight in total) included two category 1 baseball teams (Cubs and Astros), and the honorable mention teams (nine in total) included four category 1 baseball teams (Indians, Yankees, A’s, and Rays). The Red Sox, Pirates, and Cardinals are not in either list from Ferrari-King. While there are inconsistencies among these separate categorizations, there is a good deal of agreement regarding the top set of teams utilizing analytics.

Linking Adoption to Performance

Lampe (2015) conducted a similar analysis to Freeman (2016) on the 2015 MLB season and found that nearly 37% of the variance in team’s winning percentage in 2015 was explained by the team’s analytics category from ESPN (2015). He argues that most people assume that analytics usage leads to positive impacts in on-field performance, and these results provide the first glimpse of evidence that this may be true. He provides anecdotal evidence of teams with higher categorizations making the playoffs, but he also states that one year of data is not sufficient to make broader conclusions. Finally, Lampe uses results from the 2015 season and implies that 2015 is the initial year of usage; however, the ESPN rankings are based on analytics usage in 2014, thereby making 2015 the second year of analytics usage.

During this same period of time, Baumer and Zimbalist (2015) and Lindbergh and Arthur (2016) provided measures of the analytics staff size of professional baseball teams. Baumer and Zimbalist provided staff sizes for 2014 and argued that "an initial reasonable proxy for the sabermetric orientation of a team is whether or not positions are labeled analytic or sabermetric" (Baumer and Zimbalist, 2015, p. 25). Lindbergh and Arthur included staff sizes for 2009, 2012, and 2016. The correlations between these measures of staff size and the ESPN categorizations range from 0.646 to 0.762, indicating a relatively high agreement between these two measures.

LAG RESEARCH

The notion that major information technology (IT) investments by any organization will require some period of time before returns or improvements are realized was first posited nearly 30 years ago by David (1990) who attributed the delay to a necessary period of adjustment for the organization. Brynjolfsson (1993) furthered this line of thought by stating that lags are one of the possible explanations of the IT productivity paradox. Bakos (1998) referred to this lag as a diffusion delay, and this line of research was further developed by Stratopoulos and Dehning (2000), who called investments without supporting performance improvements to be irrational, and later by Goh and Kauffman (2005).

Since the mid-1990s, a great deal of research has attempted to measure this lag or diffusion delay in various industries and with various IT investments and adoptions. Mahmood, Mann, Dubrow and Skidmore (1998) argued for a two-year lag between investment in IT and improvement in financial performance; Cline and Guynes (2001) concluded that IT investment is related to firm-level performance when viewed after a two-year lag; and Feng, Chen, and Liou (2005) found productivity results for knowledge management systems implementations in the second year after implementation. Other studies have shown the lag or delay to be as high as four (Turedi and Zhu, 2012) or even six years (Yaylacicegi and Menon, 2004). Most importantly, studies of IT value, IT diffusion, and business intelligence or analytics adoption continue to incorporate a time lag or diffusion delay into their research models and continue to find support for the existence of this lag or delay (Hajli, Sims and Ibragimov, 2015; Trieu, 2017).

This isn't completely new to baseball. Lindbergh and Arthur (2016) attempted an analysis of analytics staff size on winning and found earlier adopters had greater success. Teams with an analyst in 2009 increased their winning percentage by 44 points by the 2012-14 time period (7 extra wins per season), a 3-5 year lag. Similarly, Baumer and Zimbalist (2015) noted that the Oakland A's did not have immediate success following their adoption of analytics (contrary to the implication suggested in the movie Moneyball).

HYPOTHESES, DATA COLLECTION, AND ANALYSES

Freeman (2016) suggested that one year (a single season) may not be enough time for the impact of analytics utilization to be seen in performance improvements. The IT lag research discussed earlier suggests this proposition is consistent with other IT adoptions, and a period of two or more years may be necessary before performance changes are measurable and significant. With this in mind, and considering that the original ESPN (2015) categorizations are now four years old, it is hypothesized that within four years of the original categorizations, baseball teams with higher analytics adoption categorizations will have higher winning percentages and attendance percentages than teams with lower analytics adoption categorizations. It is also hypothesized that the same effect will be seen when looking at the cumulative winning percentages across multiple seasons (as opposed to single-season winning percentages). These hypotheses are formally expressed as H1 through H3.

- H1: Teams with higher analytics categorizations will achieve higher winning percentages within four seasons.
- H2: Teams with higher analytics categorizations will achieve higher attendance percentages within four seasons.
- H3: Teams with higher analytics categorizations will achieve higher cumulative winning percentages within four seasons.

To test these hypotheses, this study uses the analytics adoption categorizations from ESPN (2015) and then uses five years of performance data from 2013-2017. For each MLB team, data from ESPN.com provided the number of wins. These data allow for the calculation of team winning percentages for each of the five years. Winning percentage is more appropriate than raw wins because sometimes, usually for weather-related reasons, a team will not play a full season. Additionally, data from ESPN.com provided the full season home attendance percentage for each team across the five seasons. As with winning percentage, attendance percentage is more appropriate than a raw attendance number as stadiums within the league have differing capacities. This percentage is the total attendance at all home games divided by the stadium's capacity for the full season (individual game capacity x home games in a season). Combining winning percentages across multiple years provides the cumulative winning percentages for 2014-2015, 2014-2016, and 2014-2017.

To maintain consistency with Freeman's (2016) data analyses, this study employed the same approaches and analyses on the previously described data regarding winning percentages, attendance percentages, and cumulative winning percentages across multiple seasons. The ESPN (2015) categorizations are based on the 2014 season. Analysis of Variance (ANOVA) tests provided the necessary comparisons of the ESPN categorizations and the performance results. The resulting p-values are shown below in Table 2. Individual cells are shaded according to significance levels of 0.05, 0.01, and 0.001 to aid in interpretation and pattern identification. In addition to the p-values, the corresponding r-squared values (coefficients of determination) are shown.

		p-value	**r-squared**
Winning %	**2014**	0.1181	0.0850
	2015	0.0004	0.3696
	2016	0.0143	0.1958
	2017	0.0181	0.1836
Attendance %	**2014**	0.4322	0.0222
	2015	0.1064	0.0904
	2016	0.0389	0.1436
	2017	0.0865	0.1013
Cumulative	**2014-15**	0.0010	0.3243
Winning %	**2014-16**	0.0006	0.3513
	2014-17	0.0000	0.4250

Table 2. ANOVA P-Values and R-Squared Values across all Variables and Years

Table 2 clearly shows significant results for analytics adoption on winning percentage. There are no significant results in 2014 (in agreement with Freeman (2016)), but for 2015, 2016, and 2017, teams see significant differences in winning percentage based on their analytics adoption categorization. The data for 2015 are consistent with Lampe (2015) who only reported the r-squared value. In terms of attendance percentage, significant results are only seen in 2016. Finally, for cumulative winning percentage, significant results are seen in all three combinations.

DISCUSSION

Hypotheses 1, 2, and 3 stated that teams with higher analytics categorizations will observe higher winning percentages, attendance percentages, and cumulative winning percentages, respectively, within four seasons. Based on the data in Table 2 and the analyses and results described in the last section, Hypothesis 1 is supported, Hypothesis 2 is only somewhat supported, and Hypothesis 3 is supported.

Interpretations and Implications

The most important finding from this research is that the winning percentages for teams with higher analytics categorizations are significantly higher in three out of the four years. This finding supports Freeman's (2016) findings for the 2014 season and supports Lampe's (2015) findings for the 2015 season by extending this previous research with data from an additional two years. This finding also quantifies the time lag of analytics adoption success in professional baseball at one year, faster than previous research in other industries but not immediate. It is interesting to note, however, that the significance levels in 2016 and 2017 are decreasing (but still significant) relative to 2015. This implies the strongest impact is in the second season and additional work is necessary to explain these decreasing significance levels in subsequent seasons. The r-squared values from these seasons demonstrate this more clearly. The r-squared in 2015 is 0.3696, while in 2016 and 2017 it falls to 0.1958 and 0.1836, respectively, meaning 37% of the winning percentages in 2015 can be explained by analytics adoption in 2014 (in agreement with Lampe) with about half of that explanatory power existing in the following two seasons.

The second point of analysis is with off-field performance as measured by attendance. Few baseball teams sellout their stadium (100% attendance) for an entire season. Most teams' attendance percentage falls between 40-80%. Significantly higher attendance percentages for teams with higher analytics categorizations were seen in only one of the four years – 2016, with an r-squared of 0.1436. While owners and general managers might argue that attendance is less important than winning, attendance impacts team revenue (tickets, concessions, and souvenirs) and creates a home-field advantage. As attendance is likely to be higher for winning teams, it is not surprising that the impact of analytics on attendance requires an additional year to see significant results. In other words, once the teams with higher analytics categorizations began to have statistically higher winning percentages in 2015, their attendance percentages became statistically higher in 2016 (the following season), though analytics adoption only explained 14% of the attendance percentages. Further research is needed to explain the non-significant findings in 2017.

The third analysis is with cumulative winning percentages over multiple years – 2014-2015, 2014-2016, and 2014-2017. The fact that all three time periods saw significant results with r-squared values between 0.32 and 0.42 indicates early adopters were able to maintain their advantage and edge over a period of time longer than a single season. The corollary is that late adopters were not able to “catch up” over time with a single season of winning.

Returning to the two research questions from the beginning of this study, performance improvements in winning percentage, attendance, and cumulative winning percentage have been found. While Freeman (2016) found no such results when looking at 2014 performance data, the inclusion of data from 2015-2017 show that lags of one year (winning percentage) and two years (attendance percentage) are observed.

Limitations

The categorizations from ESPN (2015) were subjective and may have differed somewhat if created by someone else or through a different rubric. However, there is some agreement between the ESPN categorizations and the even more subjective categories of Ferrari-King (2016). Similarly, while other measures such as staff size have been used as a proxy for analytics adoption, the ESPN categorizations go beyond staff size. Regarding the performance measures, winning percentage seems the most obvious, primary measure, but there are many others from which to choose beyond that, such as team revenue and more granular offensive or defensive statistics.

Finally, the level or intensity of analytics adoption and use in 2017 will likely be quite different than the level or intensity in 2014. This is a rapidly changing and growing field. Early adopters have likely continued to increase their adoption and usage of analytics, and early non-adopters are able to copy what the early adopters have done. However, the IT literature clearly supports the use of independent variable data from a base year to measure dependent variable data in subsequent years in order to identify the lag effect or diffusion delay.

CONCLUSION

Freeman (2016) only looked at the performance in a single year, when it is possible that the impact of analytics adoption takes longer to realize. Lampe (2015) looked at performance data in the subsequent season, but focused on the coefficient of determination and only with wins. This study extends the work of Freeman and builds on Lampe through analyses of winning percentage over time for professional baseball since 2014 and whether these performance data could have been predicted by the teams’ analytics adoption. The data also include the measurement of attendance as a secondary indicator of performance. Based on the teams’ 2014 analytics adoption as reported by ESPN (2015), analyses support the idea that statistically significant differences in teams’ winning percentages and attendance exist when looking at performance data from seasons beyond 2014 (namely, 2015-2017). In addition, the differences remain significant for multiple seasons.

Most technology implementations do not produce immediate, measurable results for the adopting organization. Time is needed for the technology to have an impact on the organization’s performance. This is no different in professional baseball. Professional baseball teams (and the MLB in general) should be aware that immediate impacts with analytics may not occur, but impacts may be realized in subsequent years.

REFERENCES

1. Alamar, B. (2013) Sports analytics: A guide for coaches, managers, and other decision makers, Columbia University Press, New York.
2. Bakos, Y. (1998) The productivity payoff of computers: A review of ‘The computer revolution: An economic perspective’ by Daniel E. Sichel, *Science*, 281, 5373, 52.
3. Baumer, B. and Zimbalist, A. (2015) The sabermetric revolution: Assessing the growth of analytics in baseball, University of Pennsylvania Press, Philadelphia.

4. Brynjolfsson, E. (1993) The productivity paradox of information technology: Review and assessment, *Communications of the ACM*, 36, 12, 67-77.
5. Cline, M. and Guynes, C. (2001) The impact of information technology investment on enterprise performance: A case study, *Information Systems Management*, 18, 4, 70-76.
6. David, P. (1990) The dynamo and the computer: A historical perspective on the modern productivity paradox, *American Economic Review*, 80, 2, 355-361.
7. ESPN. (2015) The great analytics rankings, *ESPN*, Retrieved December 10, 2018, from http://espn.go.com/espn/feature/story/_/id/12331388/the-great-analytics-rankings.
8. Eustis, S. (2018) The growing prevalence of sports analytics in 2018, *Market Research*, Retrieved December 10, 2018, from https://blog.marketresearch.com/growing-prevalence-of-sports-analytics-in-2018.
9. Feng, K., Chen, E. T. and Liou, W. (2005) Implementation of knowledge management systems and firm performance: An empirical investigation, *Journal of Computer Information Systems*, 45, 2, 92-104.
10. Ferrari-King, G. (2016) Most advanced analytics teams in sports, *Bleacher Report*, Retrieved December 18, 2018, from https://bleacherreport.com/articles/2667799-most-advanced-analytics-teams-in-sports#slide0.
11. Freeman, L. A. (2016) The impact of analytics utilization on team performance: Comparisons within and across the U.S. professional sports leagues, *Journal of International Technology and Information Management*, 25, 3, 137-160.
12. Goh, K. H. and Kauffman, R. J. (2005) Towards a theory of value latency for IT investments, in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, Big Island, HI, 231-239.
13. Hajli, M., Sims, J. M. and Ibragimov, V. (2015) Information technology (IT) productivity paradox in the 21st century, *International Journal of Productivity and Performance Management*, 64, 4, 457-478.
14. Lampe, N. (2015) 2015 playoff teams and the use of analytics, *SBNation*, Retrieved December 18, 2018, from https://www.beyondtheboxscore.com/2015/10/8/9470427/2015-playoff-teams-and-the-use-of-analytics.
15. Lindbergh, B. and Arthur, R. (2016) Statheads are the best free agent bargains in baseball, *FiveThirtyEight*, Retrieved November 28, 2018, from https://fivethirtyeight.com/features/statheads-are-the-best-free-agent-bargains-in-baseball/.
16. Mahmood, M. G., Mann, I., Dubrow, M. and Skidmore, J. (1998) Information technology investment and organization performance: A lagged data analysis, in *Proceedings of the 1998 Resources Management Association International Conference*, Idea Group Publishing, 219-225.
17. McCann, D. (2014) Predictive analytics: How clear is the ROI? *CFO Magazine*, Retrieved December 10, 2018, from http://ww2.cfo.com/technology/2014/07/predictive-analytics-clear-roi/.
18. Maxcy, J. and Drayer, J. (2014) Sports analytics: Advancing decision making through technology and data, *Institute for Business and Information Technology, Fox School of Business, Temple University*. Philadelphia, PA.
19. Seddon, P. B., Constantinidis, D. and Dod, H. (2012) How does business analytics contribute to business value, in *Proceedings of the 33rd International Conference on Information Systems*, Orlando, FL.
20. Smith, E. E. and Groetzinger, J. D. (2010) Do fans matter? The effect of attendance on the outcomes of major league baseball games, *Journal of Quantitative Analysis in Sports*, 6, 1, Article 4.
21. Stratopoulos, T. and Dehning, B. (2000) Does successful investment in information technology solve the productivity paradox? *Information & Management*, 38, 2, 103-117.
22. Trieu, V.-H. (2017) Getting value from business intelligence systems: A review and research agenda, *Decision Support Systems*, 93, 111-124.
23. Turedi, S. and Zhu, H. (2012) Business value of IT: Revisiting productivity paradox through three theoretical lenses and empirical evidence, in *Proceedings of the 18th Americas Conference on Information Systems*, Seattle, WA, 1-10.
24. Yaylacicegi, U. and Menon, N. M. (2004) Lagged impact of information technology on organizational productivity, in *Proceedings of the 10th Americas Conference on Information Systems*, New York, NY, 855-862.