Evaluating Hitting Skills of NPB Players with Logistic Regression Analysis

Mamoru Sakai[1], **Hiroki Tanioka**[2], **Kenji Matsuura**[2], **Masahiko Sano**[2], **Kenji Ohira**[2],
Tetsushi Ueta[2] **and Hiroaki Sakaguchi**[3]
[1]Graduate School of Advanced Technology and Science
Tokushima University, Japan
[2]Center for Administration of Information Technology
Tokushima University, Japan
[3]Shikoku Island League Plus, Japan

Keywords: baseball, hitting, performance, evaluation, logistic regression

1. Introduction

1.1 Background of Research

When we watch a baseball game, we often witness scenes that a line drive flies out in front of a fielder. On the other hand, a pop fly falls between infielders and outfielders. In such a scene, we viewers express “unlucky” or “lucky” for the batter. Therefore, luck is supposed to be existing in baseball. In other words, it is hard to say that hitting records depend on only the ability of the batters. In this paper, “lucky” or “unlucky” is occurred in unpredictable cases with human eyes. Hence, we consider that luck exists in a phenomenon beyond observers expected based on their observable information.

1.2 Purpose of Research

Player’s performance is appeared based on the player’s abilities affected by luck. A case study [1] said that the randomness affects player’s hitting streaky. When a player is compelled to play under unlucky situation, unfortunately, no matter how good play, it will be just treated as a failure, and vice versa. If lucky players are overestimated, they can perform under their estimated abilities in the next season. Contrariwise, if unlucky players are underestimated, they can perform over their estimated abilities in the next season. Thus, players and their team owners need an appropriate indicator of performance which produces their stable results. In other words, the indicator describes player’s authentic ability. Therefore, a purpose of this research is to clarify lucky players and unlucky players by using an indicator which is based on differences between results of two seasons in 2015 and 2016. The data including those results is provided by a workshop [2] under the sponsorship of The Institute of Statistical Mathematics.

2. Indicators of baseball players

2.1 Operations Research

Operations Research [3] is a strategy for evaluating baseball players [4]. It is one of methodology to help leaders make better dicisions using mathematical and statistical models.

2.2 Sabermetrics

There are many Key Performance Indicators (KPIs) of baseball players, such as Batting Average (AVG), the number of Home Runs (HRs), Runs Batted In (RBI), etc. These days, various methods and indicators are proposed to evaluate performance of players more objectively. One of typical indicators is Sabermetrics (Society for American Baseball Research Metrics) [5] [6]. Sabermetrics was proposed in the 1970s by Bill James who is the baseball writer. Sabermetrics is an objective method for analyzing baseball data and evaluating players. Major League Baseball (MLB) official records are based on Sabermetrics.

2.3 BABIP

There is a problem in the Sabermetrics and related researches, those which are not considering where a play is happened in the ground. BABIP (Batting Average On Ball In Play) was proposed by Voros McCracken [7][8]. BABIP is the percentage of hits on ground except Home Runs in batted ball. The BABIP equation is:

$$BABIP = \frac{H - HR}{AB - SO - HR + SF}, \tag{1}$$

where H is Hits, HR is Home Runs, AB is At Bats, SO is Strikeouts, and SF is Sacrifice Flies.

Sasaki reported the result [9] that there was low correlation between consecutive two seasons in BABIP of NPB players. According to Chris Dutton’s research using Regression Analysis [10], the batting result does not depend only on the ability of the batter, but the opponent’s defense and the ground environment. Thus, an indicator should be taken into consideration the situation on the ground to evaluate true abilities of the players.

3. Proposed Method

In this work, batting data is used as learning data, and a regression model is created using logistic regression analysis to predict a target variable. The target variable is a binary whether the batting was a hit or not. Next, a predicted BABIP as a theoretical value is calculated by the created regression model with observable information at the bat. Lastly, every luck of player is evaluated by comparing the predicted BABIP and an actual BABIP.

3.1 Logistic Regression

Logistic Regression Analysis [11] is employed for regression modeling in order to obtain hit probability of each batting. In our research, the hitting information including the situation on the ground from batting data is set as explanatory variables, and the batting result whether hit or not (1 or 0) is set as a target variable.

As explanatory variables are set on Eq. (2) as regression model, the hit probability is calculated as a value $[0, 1]$. When a batted ball except Home Runs flows into the ground, the hit probability is within the range of 0 to 1. Here, Strikeout is as 0 and Home Run is as 1. The equation of logistic regression model is given as follows.

$$p = \frac{1}{1 + \exp\left(-b_0 - \sum_{j=1}^{k} b_j x_j\right)}. \tag{2}$$

3.2 Explanatory Variable

Explanatory variables are chosen among only the information on the ground after hitting. As the ground is regarded as a lottery box, the ground information influences the distribution of lottery tickets. The following information is adopted as explanatory variable.

1) Coordinates of grounder the ball
Coordinates (x, y) of grounder, fly and line drive are converted to polar coordinates (r, θ).
2) Runner situation of each base
If a runner is on a base, the runner situation is set as 1, otherwise set as 0.
3) Defense strength of the opponent team
DER (Defense Efficiency Rating) [12] is an indicator of the team's defense strength.

It is difficult to calculate defense strength to each player. Instead, DER is employed as an explanatory variable.

$$DER = \frac{PA - H - BB - HBP - SO - E}{PA - HR - BB - HBP - SO}, \tag{3}$$

where PA is Plate Appearances, H is Hits, BB is Bases on Balls, HBP is Hit by Pitch, SO is Strikeouts, and E is Errors in the numerator. Then, PA is Plate Appearances, HR is Home Runs, BB is Bases on Balls, HBP is Hit by Pitch, and SO is Strikeouts in the denominator.

Table 1: Result of Logistic Regression Analysis.

	Estimate	std.Error	z value	p value
constant	-3.2198	0.6661	-4.833	1.34×10^{-6}
grounder r	0.0497	0.0004	102.809	2.00×10^{-16}
grounder θ	0.4937	0.0404	12.206	2.00×10^{-16}
fly r	0.0319	0.0003	90.296	2.00×10^{-16}
fly θ	0.5523	0.0346	15.946	2.00×10^{-16}
line drive r	0.0617	0.0008	76.025	2.00×10^{-16}
line drive θ	-0.4749	0.1048	-4.532	5.85×10^{-16}
First base	0.0751	0.0249	3.009	0.0026
Second base	0.0591	0.0286	2.065	0.0389
Third base	0.3216	0.0391	8.221	2.00×10^{-16}
DER	-6.0978	0.9644	-6.323	2.57×10^{-10}

* DER is Defense Efficiency Rating of opponent team.

Table 2: Statistics and Accuracies (cutoff value is 0.59)

statistics number	values	statistics measures	values
True Positive	11, 500	Sensitivity	0.57
False Positive	682	Precision	0.94
False Negative	8, 556	Specificity	0.98
True Negative	42, 761	Accuracy	0.85

3.3 Evaluation method of players in luck

The actual BABIP (ABA) is the player's practical BABIP throughout the season. The predicted BABIP (PBA) is an expectation value based on hit probabilities of a player. A luck score ($Luck$) is the actual BABIP and a difference between the predicted BABIP.

$$Luck = ABA - PBA. \tag{4}$$

It is good luck if the actual BABIP is higher than the predicted BABIP, and it is bad luck if the actual BABIP is lower.

4. Experiment

4.1 Regression Model Analysis

Table 1 is output results of the logistic regression analysis, which shows estimated coefficients and related statistics. The equation of logistic regression model in which the coefficient of the explanatory variable is substituted into the Eq. (2) is given as follows.

$$p = \frac{1}{1 + \exp\left(3.2198 - 0.0497x_1 - \cdots + 6.0978x_{10}\right)} \tag{5}$$

4.2 Prediction of batting avarages

Whether a hit probability goes to hit or not, that depend on a cutoff value. The cutoff value could be $[0, 1]$. Both true positive rate and false positive rate are moved depending on the cutoff value. The accuracy of a test is its ability to differentiate actual hitting results and predicted hitting results correctly.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6}$$

where $Accuracy$ is the accuracy, TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Table 3: Lucky players in 2015

	name	actual	predict	luck	L	R	speed
1	Atsushi Fujii	0.385	0.310	0.075	1	1	4.66
2	Takehiro Ishikawa	0.345	0.273	0.072	1	0	4.86
3	Soichiro Tateoka	0.372	0.306	0.066	1	0	5.89
4	Haruki Nishikawa	0.348	0.284	0.064	1	0	7.76
5	Shingo Kawabata	0.377	0.319	0.059	1	0	3.28
6	Takuya Nakajima	0.328	0.269	0.059	1	0	5.58
7	Ryota Imanari	0.397	0.347	0.050	1	0	2.93
8	Ryo Hijirisawa	0.346	0.297	0.049	1	0	5.73
9	Tsuyoshi Ueda	0.307	0.266	0.041	1	0	5.04
10	Kyohei Kamezawa	0.316	0.275	0.041	1	0	4.70

Table 4: Unlucky players in 2015

	name	actual	predict	luck	L	R	speed
1	Miguel Mejia	0.289	0.385	-0.096	0	1	0.99
2	Brandon J. Laird	0.244	0.310	-0.096	1	0	2.09
3	Shuichi Murata	0.264	0.316	-0.052	1	0	1.57
4	Tsubasa Aizawa	0.286	0.334	-0.048	1	0	2.79
5	Tatsuhiro Tamura	0.216	0.264	-0.048	1	0	3.94
6	Luis Cruz	0.271	0.318	-0.047	1	0	2.34
7	Keiji Obiki	0.262	0.308	-0.046	1	0	3.85
8	Mitsutaka Goto	0.252	0.298	-0.046	0	1	3.78
9	Seiichi Uchikawa	0.303	0.348	-0.045	1	0	1.96
10	Hiroyuki Nakajima	0.287	0.331	-0.044	1	0	1.88

Table 5: Lucky players with speed score in 2015

	name	actual	predict	luck	L	R	speed
1	Atsushi Fujii	0.385	0.322	0.063	1	1	4.66
2	Takehiro Ishikawa	0.345	0.287	0.058	1	0	4.86
3	Shingo Kawabata	0.377	0.320	0.057	1	0	3.28
4	Ryota Imanari	0.397	0.347	0.050	1	0	2.93
5	Soichiro Tateoka	0.372	0.329	0.043	1	0	5.89
6	Tomoya Mori	0.384	0.342	0.042	1	0	2.64
7	Takuya Nakajima	0.328	0.289	0.039	1	0	5.58
8	Hikaru Ito	0.344	0.305	0.039	0	1	2.52
9	Kazuhiro Wada	0.351	0.315	0.036	0	1	1.28
10	Akira Nakamura	0.339	0.306	0.033	1	0	3.43

Table 6: Unlucky players with speed score in 2015

	name	actual	predict	luck	L	R	speed
1	Miguel Mejia	0.289	0.357	-0.068	0	1	0.99
2	Mitsutaka Goto	0.252	0.305	-0.053	1	0	3.78
3	Brandon J. Laird	0.244	0.290	-0.046	0	1	2.09
4	Tatsuhiro Tamura	0.216	0.260	-0.044	0	1	3.94
5	Keiji Obiki	0.262	0.260	-0.042	0	1	3.85
6	Takumi Kuriyama	0.309	0.351	-0.042	1	0	2.65
7	Takahiro Okada	0.336	0.373	-0.037	1	0	3.50
8	Masahiko Morino	0.321	0.357	-0.036	1	0	1.57
9	Kazuo Matsui	0.296	0.332	-0.036	1	1	4.77
10	Ryoichi Adachi	0.261	0.297	-0.036	0	1	4.39

Table 2 shows accuracy indices. Where a cutoff value is 0.59, an accuracy is 0.85 which is the maximum value.

4.3 Lucky players and Unlucky players

Table 3 shows the top ten players whose actual BABIP was higher than the predicted BABIP in 2015. And Table 4 shows top ten players whose actual BABIP was lower than the predicted BABIP in 2015. According to the Eq. (4), actual BABIP of a lucky player is increased due to positive luck score, and actual BABIP of an unlucky player is decreased due to negative luck score. In other words, it is a player who are overestimated and underestimated in 2015. Table 3 and Table 4 include actual BABIP, predicted BABIP, and Luck. Additionally, L, R, and speed are listed. L and R mean left-handed and right-handed respectively, and speed is the speed score [13] in Eq. (7). L/R values mean that a left-handed player is 1 to L value, a right-handed player is 1 to R value, and a switch player is 1 in both L and R values. The Speed score is the following formula.

$$Spd = \frac{F1 + F2 + F3 + F4 + F5 + F6}{6}, \qquad (7)$$

where Spd means the Speed score, $F1$ means Stolen base percentage, $F2$ means Stolen base attempts, $F3$ means Triples, $F4$ means Runs scored, $F5$ means Grounded into double plays, and $F6$ means Grounded into double plays.

5. Discussion

5.1 Trend investigation

Table 3 and Table 4 show top ten lucky and unlucky players in 2015. Apparently, top ten lucky players are almost all left-handed batters and fast runners. Top ten unlucky players might be power hitters. In fact, there is a tendency that the grounder rate and the infield hit rate of the lucky players is high, but the homerun rate is low. On the other hand, there is another tendency that the fly rate of the unlucky players is high, but the infield hit rate tends is low. There is a similar tendency in top ten lucky and unlucky players in 2016.

5.2 PCA (Principal Component Analysis)

Trend between lucky players and unlucky players can be invested using PCA (Principal Component Analysis) with batting statistics. Figure 1 show distributions with PC1, PC2, and PC3 of the principal components. From the PCA results, it seems that there is a difference between lucky players and unlucky players, which is derived from player's running ability. This fact must be not overlooked. Luck must be the influence by information that an observer cannot observe. However, we can know the player is left-handed and has high running ability.

5.3 Distributions of Luck

From the trend investigation, lucky players have left-handed and high running ability. Hence, some conclusive factors must be contained in information used in regression analysis. Then, two logistic regression analyses with L/R values and the speed score are tested for calculating predicted BABIP.

Table 5 and Table 6 show top ten lucky players and unlucky players in 2015. There are differences compared to Table 3 and Table 4. From the differences, the bias between lucky players and unlucky players looks smaller in the L/R values and the speed scores. In Table 3 and Table 4, the average speed score of the lucky players is 5.04, and the average of the unlucky players is 2.51. However, in Table 5 and Table 6, the average speed score is 3.71 and 3.15. Table 7

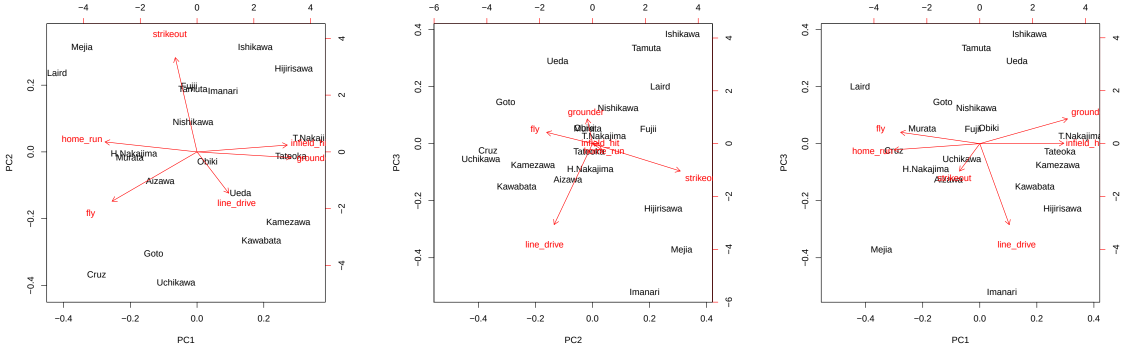

Fig. 1: Distributions with PC1, PC2, PC3 as axes

Table 7: Luck distributions in two seasons 2015–2016.

	mean	variance	players	df	p value
Luck Dist	-9.57×10^{-5}	0.000964	209	208	-
Luck Dist L/R	0.000282	0.000877	209	208	0.247
Luck Dist Spd	0.000617	0.000734	209	208	**0.0249**

shows statistical information including variance values of the luck score distributions in 2015–2016.

There is not a statistically significant reduction 0.000087 between Luck Dist and Luck Dist L/R in variance. There is a statistically significant reduction 0.000230 between Luck Dist and Luck Dist Spd in variance. Here, p is $0.0249 (< 0.05)$ in F-test. From these results, the predicted BABIP got closer to the actual BABIP, while the luck score distribution was shrunk. This means that the more observable information is adopted to logistic regression as explanatory variable, the more the luck score distribution is shrunk. If all the information on the ground is observable, luck is not existing in baseball games.

6. Conclusion

Investigating the tendency of lucky players and unlucky players, lucky players hit many grounders and could be fast runners, unlucky players hit many fly balls and might be slow runners. It seems that there are still factors to be considered such as player's ability in the part which was influenced by luck. When the L/R values and the speed score were adopted to logistic regression as explanatory variables, the bias derived from player's ability was slightly decreased in the luck score.

Acknowledgment

This research was supported by Data Stadium Inc. and The Institute of Statistical Mathematics. We thank our colleagues from A6 Computer System Laboratory of Tokushima University who provided insight and expertise that greatly assisted the research, although they may not agree with all the conclusions of this paper. We thank Michitomo Morii for assistance with particular technique, and Kohei Kawanaka for comments that greatly improved the manuscript.

It was found from the result that the predicted BABIP properly evaluated player's hitting ability rather than the actual BABIP, because the predicted BABIP has smaller influence of luck than the actual BABIP. However, it is not enough to gathering observable information on the ground, because, a speed score is a pseudo speed score which does not represent a player's running ability to the first base every at bat. Therefore, we need to measure every arrival time for reaching the first base after hitting.

References

[1] S. C. Albright, "A Statistical Analysis of Hitting Streaks in Baseball," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1175–1183, 1993.

[2] M. Sakai and H. Tanioka, "Yakyu ni Okeru Un toha Nanika - Logistic Kaiki Bunseki wo Mochiita Anda Kakuritsu no Yosoku (What is Lucy in Baseball? – Prediction of Hit Probability Using Logistic Regression)," The Institute of Statistical Mathmatics, Tech. Rep., March, published in Japanese, 2018.

[3] C. W. Churchman, R. L. Ackoff, Ackoff, and E. Arnoff, "Introduction to operations research," 1957.

[4] N. Streib, S. J. Young, and J. Sokol, "A Major League Baseball Team Uses Operations Research to Improve Draft Preparation," vol. 42, pp. 119–130, March 2012.

[5] M. M. Lewis., "Moneyball: The Art of Winning an Unfair Game," NY, USA, 2003.

[6] R. J. Puerzer, "From Scientific Baseball to Sabermetrics: Professional Baseball as a Reflection of Engineering and Management in Society," *NINE: A Journal of Baseball History and Culture*, vol. 11, no. 1, pp. 34–48, 2002, accessed: 2018-02-14.

[7] V. McCracken, "Pitching and Defense: How Much Control Do Hurlers Have?" https://www.baseballprospectus.com/news/article/878/pitching-and-defense-how-much-control-do-hurlers-have/, January 2001, accessed: 2018-02-14.

[8] D. Studeman, "Data Erratum Et Cetera," https://www.fangraphs.com/tht/data-erratum-etcetera/, January 2004, accessed: 2018-02-14.

[9] H. Sasaki, "BABIP ga Imisuru Tokoro to, sono Kaishaku no Muzukashisa (The meaning of BABIP and the Difficulty of the Interpretation)," http://www.baseball-lab.jp/column/entry/175/, May, published in Japanese, 2015, Accessed: 2018-02-14.

[10] C. Dutton, "Batters and BABIP," https://www.fangraphs.com/tht/batters-and-babip/, December 2008, accessed: 2018-02-14.

[11] S. R. Bailey, "Forecasting Batting Averages in MLB," Master's thesis, Simon Fraser University, November 2017.

[12] "Defensive Efficiency Ratio (DER)," http://m.mlb.com/glossary/advanced-stats/defensive-efficiency-ratio, accessed: 2018-02-14.

[13] B. James, *The Bill James Baseball Abstract 1987 (1st ed.)*. Ballantine Books, 1987, accessed: 2018-02-14.