Evaluation of Major League Baseball Offensive Statistics Underscores Shohei Ohtani's Exceptional Batting Performance

Noriaki Endo, *Member, IAENG,* Yasumasa Yamaguchi, Hidetaka Uchino, and Chiaki Hashimoto

***Abstract*—This study aims to objectively evaluate the offensive performance of Major League Baseball players. We analyzed offensive statistics for the 2023 season, focusing on the top 30 players in the American League by batting average. Principal component analysis was conducted to reduce the number of parameters and summarize key aspects of batting performance. The first and second principal components effectively captured players' overall abilities. Scatter plots of these components also clearly illustrated distinct player characteristics. Notably, Shohei Ohtani's scores differed markedly from those of other players, indicating his exceptional performance even when assessed solely on 2023 batting data.**

***Index Terms*—batting performance; Major League Baseball; offensive statistics; principal component analysis (PCA); Shohei Ohtani**

I. Introduction

Major League Baseball fans often discuss player statistics as well as the game results of their favorite teams. A player's offensive performance can be assessed using many parameters, such as batting average, home runs, and on-base plus slugging (OPS). Therefore, it can be difficult to grasp a batter's overall ability at a glance. This study aims to objectively evaluate the offensive performance of Major League Baseball players. To this end, we conduct a multivariate analysis to develop statistical measures of players' overall abilities. As the primary statistical method, principal component analysis (PCA) is employed to reduce the number of parameters and summarize key aspects of batting performance. Additionally, players are grouped according to offensive performance profiles.

II. Research Methods

A. Research Materials

We analyzed MLB players' offensive statistics for the 2023 season, obtained from the Yahoo Japan website [1] and summarized in Table I. It consists of data on the 30 players with the highest batting averages in the MLB's American League.

Manuscript received February 27, 2025; revised May 16, 2025.

N. Endo is a professor in the Department of Sports Intelligence and Mass Media, Faculty of Sports Science, Sendai University, Shibata, Miyagi Prefecture, Japan (corresponding author, Email: endo@iwate-u.ac.jp).

Y. Yamaguchi is an assistant professor in the Department of Sports Intelligence and Mass Media, Faculty of Sports Science, Sendai University, Shibata, Miyagi Prefecture, Japan (Email: ys-yamaguchi@sendai-u.ac.jp).

H. Uchino is an associate professor in the Department of Sports Intelligence and Mass Media, Faculty of Sports Science, Sendai University, Shibata, Miyagi Prefecture, Japan (Email: hd-uchino@sendai-u.ac.jp).

C. Hashimoto is an associate professor in the Department of Sports Intelligence and Mass Media, Faculty of Sports Science, Sendai University, Shibata, Miyagi Prefecture, Japan (Email: ca-hashimoto@sendai-u.ac.jp).

B. Parameters Used in this Research

All parameters used in this study are classified as Standard Stats by Major League Baseball [2]. These include batting average (AVG), triples (3B), home runs (HR), total bases (TB), runs batted in (RBI), runs (R), strikeouts (SO or K), walks (BB), stolen bases (SB), ground into double plays (GIDP), on-base percentage (OBP), slugging percentage (SLG), on-base plus slugging (OPS), and batting average with runners in scoring position.

C. System

1) Hardware: The client computer system was a standard PC laptop, an ASUS Expert Book with an Intel Core i5-10210U CPU @ 1.60 GHz, 8 GB memory, and a 512 GB SSD.

2) Software: For basic calculations, we used LibreOffice Calc 7.3.6.2 (x64) spreadsheet software [3] running on the Windows 10 Professional operating system. For statistical analysis, we used R System version 4.2.2 [4] running on the same operating system. Both are well-known open-source applications used worldwide.

D. Data Processing

Using the acquired data, we conducted a PCA [5] with the R statistical system [4], employing the princomp function. PCA summarizes multiple variables into a smaller number of components, making the data easier to interpret. This method has been used in several studies related to Major League Baseball [6], [7], [8]. Furthermore, Attarian et al. [9],[10] placed a particular emphasis on improving the accuracy of Bayesian classifiers [11] through feature selection and dimension reduction via linear discriminant analysis (LDA) and PCA, respectively.

III. Results and Discussion

A. Results of PCA

1) Table II summarizes the PCA, focusing on the first five components whose standard deviations exceed 1.00—specifically, 2.370, 1.475, 1.292, 1.251, and 1.030. The proportions of variance explained by components 1 through 5 were 0.401, 0.155, 0.119, 0.112, and 0.076, respectively. The cumulative proportions of variance explained by these components were 0.401, 0.556, 0.676, 0.787, and 0.863, respectively.

2) Figure 1 illustrates the coefficients (eigenvectors) of the first and second principal components, calculated using PCA.

3) Figure 2 displays a scatter plot of the scores for the first and second principal components, also obtained through PCA.

TABLE I
OFFENSIVE STATISTICS FOR THE 2023 SEASON, FOCUSING ON THE TOP 30 PLAYERS IN THE AMERICAN LEAGUE BY BATTING AVERAGE. "BA W/RISP" MEANS "BATTING AVERAGE WITH RUNNERS IN SCORING POSITION".

	BA	Triple	HR	Total Base	RBI	Run	SO	Walk	SB	GIDP	OBP	SLG	OPS	BA w/RISP
Diaz(TB)	0.33	0	22	274	78	95	94	65	0	16	0.41	0.522	0.932	0.365
Seeger(TEX)	0.327	0	33	297	96	88	88	49	2	9	0.39	0.623	1.013	0.385
Bichette(TOR)	0.306	3	20	271	73	69	115	27	5	14	0.339	0.475	0.814	0.364
Ohtani(LAA)	0.304	8	44	325	95	102	143	91	20	9	0.412	0.654	1.066	0.317
Yoshida(BOS)	0.289	3	15	239	72	71	81	34	8	20	0.338	0.445	0.783	0.266
Tucker(HOU)	0.284	5	29	297	112	97	92	80	30	11	0.369	0.517	0.886	0.354
Ramirez(CLE)	0.282	5	24	290	80	87	73	73	28	8	0.356	0.475	0.831	0.212
Latchman(BAL)	0.277	1	20	256	80	84	101	92	1	14	0.374	0.435	0.809	0.274
Witt Jr.(KC)	0.27613	11	30	317	96	97	121	40	49	11	0.319	0.495	0.813	0.29
Semien(TEX)	0.27612	4	29	320	100	122	110	72	14	5	0.348	0.478	0.826	0.322
J.Turner(BOS)	0.276	0	23	254	96	86	110	51	4	10	0.345	0.455	0.8	0.338
Rodriguez(SEA)	0.2752	2	32	317	103	102	175	47	37	14	0.333	0.485	0.818	0.299
Hayes(BAL)	0.275	2	16	231	67	76	141	38	5	11	0.325	0.444	0.769	0.271
Torres(NYY)	0.273	2	25	270	68	90	98	67	13	19	0.347	0.453	0.8	0.27
Merrifield(TOR)	0.2724	0	11	209	67	66	101	36	26	15	0.318	0.382	0.7	0.296
Garcia(KC)	0.2716	4	4	166	50	59	115	38	23	9	0.323	0.358	0.681	0.321
Devers(BOS)	0.271	0	33	290	100	90	126	62	5	15	0.351	0.5	0.851	0.289
Kwan(CLE)	0.268	7	5	236	54	93	75	70	21	9	0.34	0.37	0.71	0.292
Taveras(TEX)	0.2661	3	14	215	67	67	117	35	14	3	0.312	0.421	0.733	0.237
Crawford(SEA)	0.2659	0	19	234	65	94	125	94	2	7	0.38	0.438	0.818	0.263
Yang(TEX)	0.2657	1	23	223	70	75	151	30	1	7	0.315	0.467	0.781	0.252
Guerrero Jr.(TOR)	0.2641	0	26	267	94	78	100	67	5	23	0.345	0.444	0.788	0.268
Verdugo(BOS)	0.2637	5	13	230	54	81	93	45	5	11	0.324	0.421	0.745	0.25
Robert(CWS)	0.2637	1	38	296	80	90	172	30	20	10	0.315	0.542	0.857	0.212
Pena(HOU)	0.26343	3	10	220	52	81	129	43	13	11	0.324	0.381	0.705	0.258
Casaz(BOS)	0.2634	2	24	210	65	66	126	70	0	7	0.367	0.49	0.856	0.25
Bregman(HOU)	0.2621	4	25	274	98	103	87	92	3	22	0.363	0.441	0.804	0.285
Drury(LAA)	0.2619	3	26	241	83	61	136	25	0	14	0.306	0.497	0.803	0.281
Low(TEX)	0.26164	3	17	258	82	89	165	93	1	22	0.36	0.414	0.775	0.264
Benintendi(CWS)	0.26157	2	5	200	45	72	89	52	13	12	0.326	0.356	0.682	0.333

TABLE II
PRINCIPAL COMPONENT ANALYSIS RESULTS

	comp.1	comp.2	comp.3	comp.4	comp.5
Standard Deviation	2.370	1.475	1.292	1.251	1.030
Proportion of Variance	0.401	0.155	0.119	0.112	0.076
Cumulative Proportion	0.401	0.556	0.676	0.787	0.863

TABLE III
COEFFICIENTS FOR PRINCIPAL COMPONENTS. "BA W/RISP" MEANS "BATTING AVERAGE WITH RUNNERS IN SCORING POSITION".

	comp.1	comp.2	comp.3	comp.4	comp.5
Batting Average (AVG)	0.263	0.307	0.006	0.441	0.066
Triple (3B)	0.054	-0.372	-0.461	0.199	-0.099
Home Run (HR)	0.363	-0.202	0.261	-0.086	0.013
Total Bases (TB)	0.374	-0.216	-0.050	-0.055	0.187
Runs Batted In (RBI)	0.341	-0.139	0.055	-0.125	0.328
Run (R)	0.299	-0.151	-0.294	-0.224	-0.038
Strikeout (SO, K)	0.030	-0.332	0.492	-0.161	-0.099
Walk (BB)	0.187	0.168	-0.378	-0.487	-0.283
Stolen Base (SB)	0.040	-0.479	-0.330	0.266	0.162
Ground Into Double Play (GIDP)	0.016	0.184	-0.027	-0.384	0.720
On-base Percentage (OBP)	0.306	0.361	-0.162	-0.113	-0.264
Slugging Percentage (SLG)	0.371	-0.041	0.261	0.158	-0.125
On-base Plus Slugging (OPS)	0.392	0.086	0.152	0.087	-0.184
BA w/RISP	0.168	0.310	-0.124	0.404	0.297

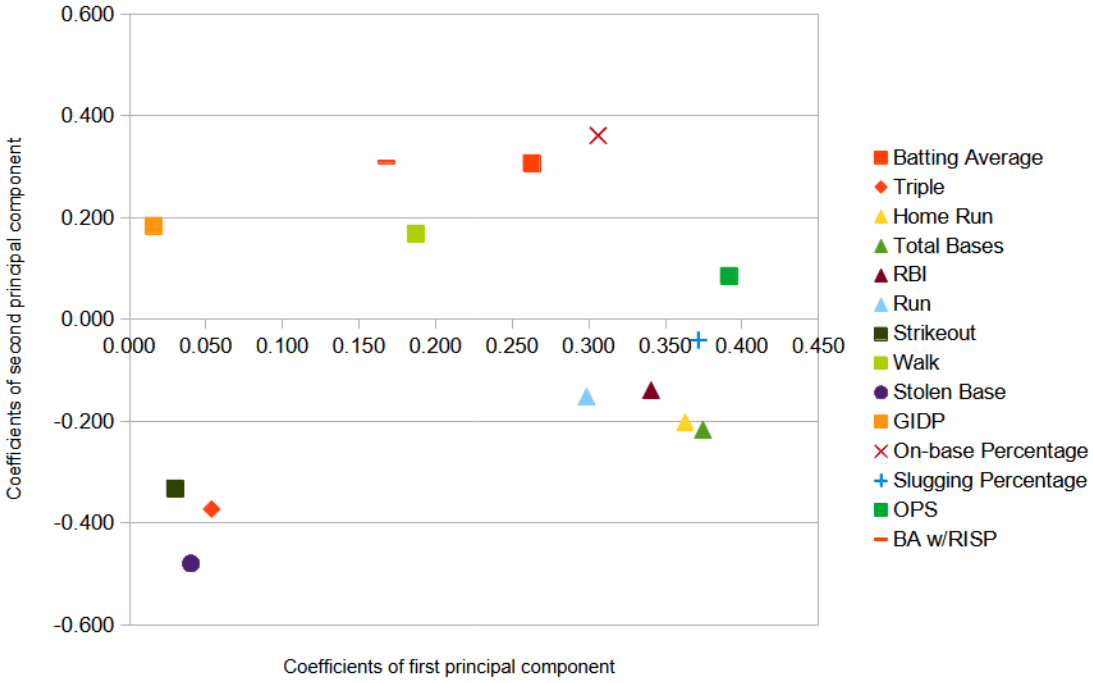

Fig. 1. Coefficients of first and second principal components calculated by principal component analysis

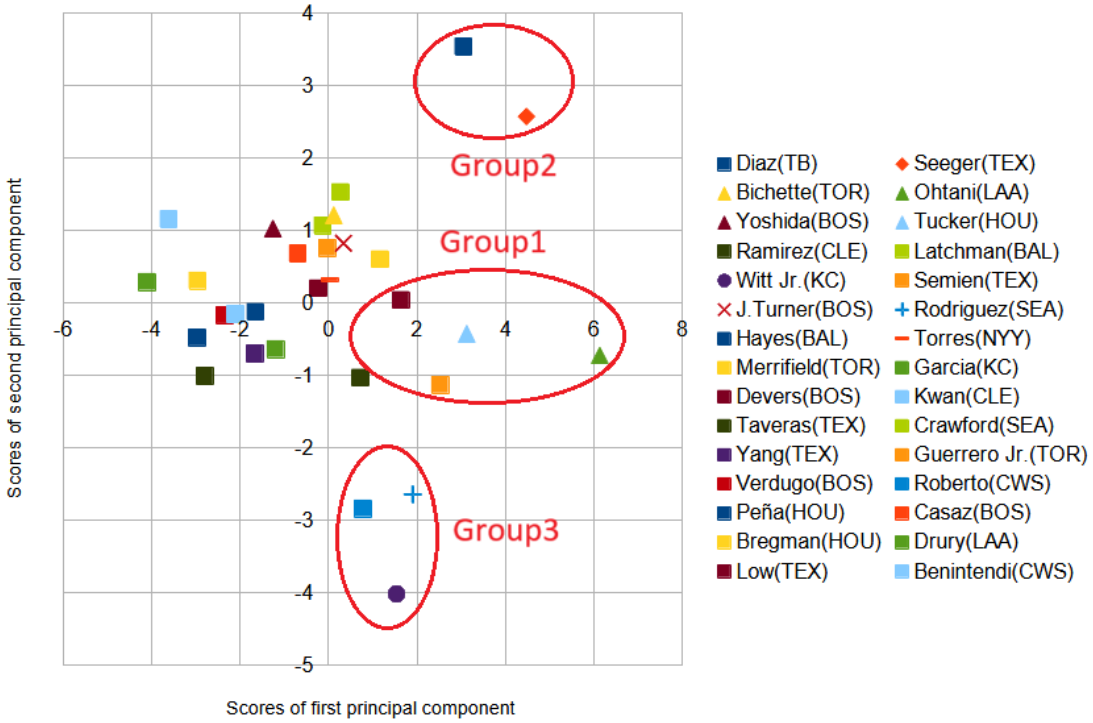

Fig. 2. Scatter plot of the scores of first and second principal components calculated by principal component analysis

B. Interpretation of Principal Component Analysis

The present results have the following implications:

1) For the first principal component, the coefficients of OPS (0.392), HR (0.363), and RBI (0.341) are high. Therefore, the first principal component can be regarded as representing players' power-hitting ability.

2) For the second principal component, the coefficients of on-base percentage (0.361) and batting average (0.307) are high, while those of stolen bases (-0.479) and triples (-0.372) are low. Therefore, the second principal component can be regarded as representing players' contact hitting and speed abilities.

C. Classification of Players into Characteristic Groups

We can identify characteristic groups in Fig. 2.

1) Group One

Group One consists of players with a high first principal component and a midrange second principal component. This group includes Ohtani (LAA), Tucker (HOU), Semien (TEX), Devers (BOS). It is characterized by high OPS, high RBI, and high HR, indicating that these players are power hitters.

2) Group Two

Group Two consists of players with both a high first principal component and a high second principal component. This group includes Diaz (TB) and Seager (TEX). It is characterized by a high on-base percentage and a high batting average, indicating contact hitters with strong consistency at the plate.

3) Group Three

Group Three consists of players with a midrange first principal component and a low second principal component. This group includes Witt Jr. (KC), Robert (CWS), and Rodriguez (SEA). It is characterized by a high number of stolen bases and a high strikeout rate, suggesting speed-oriented players with high strikeout rates.

D. These results underscore Shohei Ohtani's exceptional batting performance

As Figure 2 shows, Ohtani's score for the first principal component is extremely high, while that for the second is midrange. His scores are plotted in clearly distinct positions relative to other players. These results underscore Ohtani's exceptional performance, even when evaluated solely based on 2023 batting statistics.

IV. Conclusions

This study applied PCA to condense multiple batting statistics into a smaller set of key performance indicators. We found that the first and second principal components effectively capture players' overall abilities. Additionally, scatter plots of these components clearly illustrate differences in player characteristics. Notably, Shohei Ohtani's scores were plotted in distinct positions compared to other players. These results highlight Ohtani's exceptional performance, even when evaluated solely based on 2023 batting statistics.

References

[1] Yahoo Japan MLB (accessed on January 1, 2024) https://baseball.yahoo.co.jp/mlb/

[2] Official Site of Major League Baseball, Offence Category of Standard Stats (accessed on January 1, 2024) https://www.mlb.com/glossary/standard-stats

[3] Official Website of the LibreOffice Project (accessed on January 1, 2024) http://www.libreoffice.org/

[4] The R Project for Statistical Computing (accessed on January 1, 2024) https://www.r-project.org/

[5] Greenacre, M., Groenen, P.J.F., Hastie, T. et al. Principal component analysis. Nat Rev Methods Primers 2, Article number 100, 2022. Available at https://doi.org/10.1038/s43586-022-00184-w

[6] Depken, C.A., Grant, D., "Multiproduct pricing in Major League Baseball: A principal components analysis," *Economic Inquiry*, vol.49, no.2, pp474-488, 2011.

[7] Gushiken, S., Ikezaki, J., Miyata, R., "Principal component analysis of starting pitcher indexes in Nippon professional baseball," *ICIIBMS 2015 - International Conference on Intelligent Informatics and Biomedical Sciences*, pp378-379, 7439490, 2016.

[8] Matsuka, H., Asahi, Y., "What kind of foreign baseball players want to get Japanese baseball team?," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9735, pp560-568, 2016.

[9] Adam Attarian, George Danis, Jessica Gronsbell, Gerard Iervolino, and Hien Tran, "A Comparison of Feature Selection and Classification Algorithms in Identifying Baseball Pitches," *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2013*, IMECS 2013, 13-15 March, 2013, Hong Kong, pp263-268.

[10] Attarian A., Danis G., Gronsbell J., Iervolino G., Layne L., Padgett D., Tran H., "Baseball pitch classification: A Bayesian method and dimension reduction investigation," *IAENG Transactions on Engineering Sciences - Special Issue of the International MultiConference of Engineers and Computer Scientists 2013 and World Congress on Engineering 2013*, IMECS 2013 and WCE 2013, Routledge: Taylor & Francis Group, pp393-399, 2014.

[11] Lei Jiang, Peng Yuan, Qiongbing Zhang, and Qi Liu, "A Study of the Naive Bayes Classification Based on the Laplacian Matrix," *IAENG International Journal of Computer Science*, vol.47, no.4, pp713-722, 2020.

Appendix

All parameters used in this research are categorized as Standard Stats by Major League Baseball [2].

1) Batting Average (AVG): One of the oldest and most widely used measures of a hitter's success at the plate. It is calculated by dividing a player's total hits by his total at-bats, yielding a number between .000 and 1.000. In recent years, the league-wide batting average has typically hovered around .250.

2) Triple (3B): Often called "the most exciting play in baseball," a triple occurs when a batter hits the ball into play and reaches third base without the aid of an error or an attempt to retire another baserunner.

3) Home Run (HR): A home run occurs when a batter hits a fair ball and scores on the play without being put out and without the benefit of an error.

4) Total Bases (TB): Total bases refer to the number of bases a batter gains through hits. A single counts as one base, a double as two, a triple as three, and a home run as four total bases.

5) Runs Batted In (RBI): A batter is credited with an RBI in most cases where his plate appearance results in a run being scored. However, a player does not receive an RBI if the run scores due to a fielding error or when grounding into a double play.

6) Run (R): A player is awarded a run if he safely crosses home plate. The manner in which the player reached base

is not considered; for example, reaching home by error or fielder's choice still results in a run. A pinch-runner who scores is also credited with a run.

7) Strikeout (SO, K): A strikeout occurs when a batter accumulates three strikes—either swinging or called—during a plate appearance. A foul ball counts as a strike unless it would be the third strike, in which case it does not. However, a foul tip caught by the catcher is considered a third strike.

8) Walk (BB): A walk (or base on balls) occurs when a pitcher throws four pitches outside the strike zone that are not swung at by the batter. The batter is then awarded first base. In the scorebook, a walk is denoted by the letters BB.

9) Stolen Base (SB): A stolen base occurs when a baserunner advances to a base he is not entitled to, typically while the pitcher is delivering the ball to home plate. It can also occur while the pitcher still holds the ball, during a pickoff attempt, or as the catcher is returning the ball to the pitcher.

10) Ground Into Double Play (GIDP): A GIDP occurs when a player hits a ground ball that results in multiple outs on the bases. The most common type involves a forceout on the runner advancing from first to second base, followed by a second forceout on the batter running to first.

11) On-base Percentage (OBP): OBP refers to how frequently a batter reaches base per plate appearance. Times on base include hits, walks, and hit-by-pitches, but do not include errors, fielder's choices, or dropped third strikes. (Sacrifice bunts are excluded from the equation entirely, as it is rarely a hitter's decision to sacrifice himself; rather, it is typically a manager's choice as part of in-game strategy.)

12) Slugging Percentage (SLG): Slugging percentage represents the total number of bases a player records per at-bat. Unlike on-base percentage, slugging percentage focuses solely on hits and does not include walks or hit-by-pitches in its calculation.

13) On-base Plus Slugging (OPS): OPS combines on-base percentage and slugging percentage into a single metric. It reflects both a hitter's ability to reach base and his capacity to hit for average and power.

14) Batting Average with Runners in Scoring Position (BA w/RISP): This statistic is calculated by dividing a player's hits by his total at-bats when there are runners in scoring position.