Factor Analysis of the Batting Average

Hiroki Yamato[1(✉)] and Yumi Asahi[2]

[1] School of Information and Communication Studies Department of Management Systems Engineering, Tokai University, Tokyo, Japan
5bjmll24@mail.u-tokai.ac.jp

[2] School of Information and Telecommunication Engineering, Department of Management System Engineering, Tokai University, Tokyo, Japan
asahi@tsc.u-tokai.ac.jp

Abstract. This study is factor analysis of the batting average in the professional baseball in Japan. We analyze the factor influencing the batting average using the Japanese professional baseball data. There is no established method to ensure a good shot at Japanese baseball. Based on the results, clarify factors that prevent pitchers from hitting hits and factors that batters increase hits. And establish baseball teaching methods based on the clarified factor. Finally, we aim to improve the level of the professional baseball world of Japan.

The data used are the one-ball data in the regular season of Japanese professional baseball in 2015 and 2016. One-ball data is data every time a pitcher throws one ball to a batter. This time, we used only the data of the battle of right-handed pitcher and right-handed batter. The reason for limiting the data is that it is judged that it is easier to extract the characteristics of the factor when narrowing down the conditions.

In this research, factor analysis is performed first, and covariance structure analysis is performed based on extracted factors. Factor analysis extracts pitcher and batters how to approach the ball. In the covariance structure analysis, we analyze how the extracted factor affects variables.

The result of the factor analysis is that the pitcher can extract four factors, the batter can extract two factors. We named the extracted pitcher's factors "throw down low", "throw falling balls", "throw balls to escape outside", "attack in-course". We named the extracted batter's factors "upper swing", "down swing". When covariance structure analysis was performed using the result of the factor analysis, three models could be created. The three models can know how each factor influences hits, outs, batting average. From the results of these models, upper swing had a positive influence on hits, and it turned out that it had a bad influence on outs. It also proved to have a positive effect on latent variable batting time consisting of hits and outs. In summary, it turns out that doing an upper swing has a good influence on increasing the batting average.

From the analysis result, it turned out that the upper swing is important for improving the batting average. The future task can be to analyze also combinations other than right-handed pitcher versus right-handed batter who could not be done this time. In addition, we clarify the explanatory variable which has the most influence on improving batting average among latent variable upper swing.

Keywords: Sports marketing · Factor analysis

S. Yamamoto and H. Mori (Eds.): HIMI 2018, LNCS 10904, pp. 145–155, 2018.
https://doi.org/10.1007/978-3-319-92043-6_12

1 Introduction

Baseball is one of the ball games that hits a ball with a bat. It is a popular sport in Japan, the USA, Cuba etc. especially major league in America is considered to be the best league of the baseball world. The reason is that the league's economic scale is the biggest in the world. According to Fig. 1, we can see the average salary of MLB players is about 11 times that of NPB players in 2015. MLB players can get high salary. Therefore, it is easy for players to gather from all over the world and the number of player increase. Actually, NPB has 11 teams and MLB holds 30 teams. As a result, the competition of higher level is born, and powerful players are born. From Japan, sending out major players who can succeed with major like Ichiro. Japanese baseball league which Ichiro is belongs to is also high level. Besides Ichiro, Japanese baseball league has sent top players such as Hiroki Kuroda and Hideki Matsui to the major league. However, the level of Japanese professional league is starting to be thought to be getting lower in recent years. The reason is that the number of athletes trying to challenge Major League every year is decreasing from Fig. 2 and the Japanese major league challenge record that eventually lasted 22 years broke. Pitchers are active with Yu Darvish and Masahiro Tanaka and others, but no one is active in the fielder. Fielders from Japan may be disappeared in major league this year. It was third in WBC in 2017, but it seems to be struggling compared to when we were able to win successive. Therefore, raising the level of professional baseball in Japan is an urgent task and we examined the method for doing so. The problem of Japanese baseball world is various instruct method. There are too many methods, which is confusing. For example, strike to a ball from above, strongly swing the bat, etc. but it is a way to increase strong hitting and not a way to increase hits. In Japan, the method to increase hits is not clarified now. Therefore, this study is clarified factor of the increasing hits and batting average in the professional baseball in Japan and propose an instruct method to increase batting average and number of hits to improve batting record.

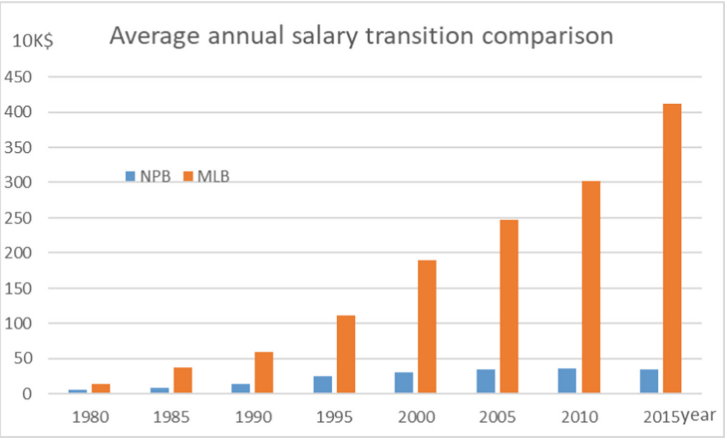

Fig. 1. Average annual salary transition comparison

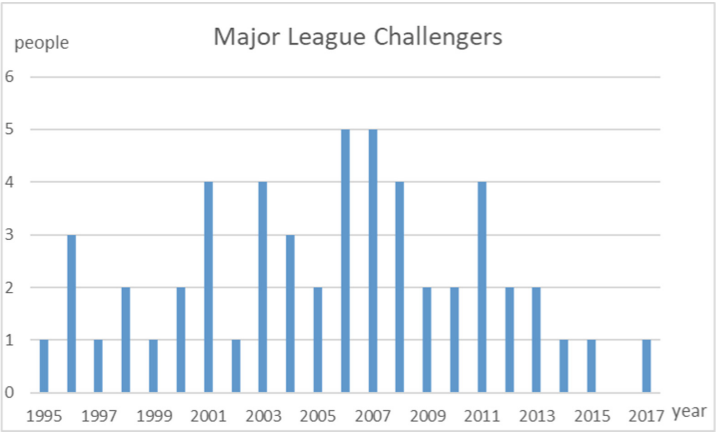

Fig. 2. Number of major league challengers

2 Detail of Used Data

We used the one-ball data in the regular season of Japanese professional baseball in 2015 and 2016. It was provided by Data Stadium. One-ball data is data every time a pitcher throws one ball to a batter. This time, used variables in this data are summarized in Table 1 below.

Table 1. Detail of used data

Details of data	
Flying distanceX	Numerical data from 0 to 280
Flying distanceY	Numerical data from 0 to 280
Hitting segment	1 when it was fly, 2 when it was a grounder, 3 when it was a liner, 4 if batter missed it
Swing	1 or 0 qualitative data
Pitching courseX	Numerical data from 0 to 200
Pitching courseY	Numerical data from 0 to 250
Breaking ball	1 for a curve, 2 “for a cutter, 3 for a shoot 4 for a shinker, 5 “for a straight 6 for a slider, 7 for a split 8 for a Off-speed pitch, 9 for a special ball
One-bound ball	1 or 0 qualitative data
Ball speed	Numerical data from 82 to 165
Batting result classification	1 is when it was swing and miss, 2 is when it was looking, 3 is when it was hit 4 is when it was out 5 is when it was home run, 6 is when it was foul, 7 is when it was bunt

Flying distance coordinates X and Y in Fig. 3 are the upper left corner of the figure above as the origin. For example, in the case of a catcher fly, the distance coordinate X

is 40 and the distance of the distance Y is 240. Swing is that whether the batter swung or not. Hitting segment is batting segment. For example, fly or grounder or liner. The pitching coordinates X and Y in Fig. 4 represent the height as the Y -axis and width of the zone as the X-axis. Breaking ball is what kind balls threw. For example, slider, off-speed pitch, etc. One-bound ball is that whether the ball has bounced one or not. Ball speed is that speed of the thrown ball. Flying distanceX, flying distanceY, pitching courseX, pitching courseY and ball speed are quantity data. Swing, Result of batting, one-bound ball and breaking ball are qualitative data. The scale of them is different. Therefore, we need to align the scales before starting the analysis. This time, we try to do factor analysis and covariance structure analysis. The purpose of factor analysis is to extract elements of batter and pitcher. At the same time as converting qualitative data to quantitative data, we convert it so that features can be easily grasped by factor analysis. Details of conversion are summarized in Table 2.

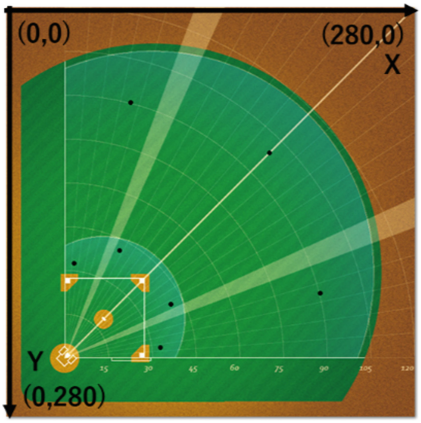

Fig. 3. Flying distance coordinates

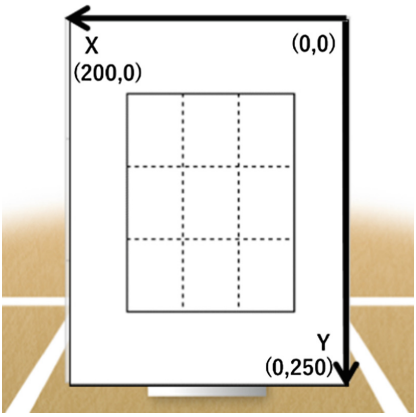

Fig. 4. Pitching coordinates

Flying distanceX and Flying distanceY are very hard to use. Therefore, we redefine flying distanceX2 and flying distanceY2 from them. They adjusted so that the origin overlaps the home base. We drew a line by dividing places other than the foul area into three at 30 degrees each like Fig. 5. These lines are the line of $Y = \sqrt{3}x$ and line of $Y = 1/\sqrt{3}x$. ① which the place surrounded by Y-axis and the line of $Y = Y\sqrt{3}x$ is named left direction. ② which the place surrounded by the line of X-axis and the line of $Y = 1/\sqrt{3}x$ is named right direction. When it is caught in these zones, it is set to 1, and when it is not caught 0. In this manner, the coordinate data is converted into qualitative data. We also redefine variable distance in this data. The distance data can be obtained using the by how far the origin is from these flying distance X2 and flying distance Y2. The formula is distance $= \sqrt{(\text{flying distance x2\^{}2} + \text{flying distance y2\^{}2})}$. We redefine short distance and long distance from this distance. First, draw a line to classify the distance. The black line in the figure is a line representing the part where the distance from the origin is 42, the red line is the line showing the distance 162 like

Fig. 6. We redefine the inner side of the black line as a short distance and the outer side of the red line as a long distance. Set it to 1 when the hit ball flew into the defined place and 0 otherwise. From batting tendency, we redefine fly tendency or grounder tendency. Fly tendency is set to 1 when the ball fly and 0 otherwise. Grounder tendency is 1 when the ball rolls, and 0 otherwise.

Table 2. Change of variable

The original variable	New variable
Flying distanceX, flying distanceY	Flying distanceX2, flying distanceY2
Flying distanceX2, flying distanceY2	Left direction, right direction, distance
Distance	Short distance, long distance
Hitting segment	Fly, grounder
Batting result classification	Hitout
Pitching courseX, pitching courseY	High, low, inside, outside
Speed of ball	Fast ball, slow ball
Breaking ball	Bending ball, Falling ball, shinker

Next, we redefine pitching coordinates X and Y as inside, outside, higher and lower. In order to classify inside and outside, the zone was divided into three zones and a line was drawn like Fig. 7. It drew a line so that the proportion of thrown balls was equal for each zone. Similarly, classify high and low line to be drawn like Fig. 8. Inside is 1 when the ball is thrown into ① in Fig. 1, and 0 otherwise. Outside is 1 when the ball is thrown into ② in Fig. 1, and 0 otherwise. Higher is 1 when the ball is thrown into ① in Fig. 2, and 0 otherwise. Lower is 1 when the ball thrown into ② in Fig. 2,

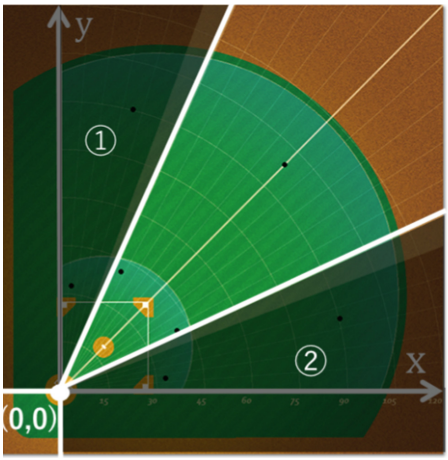

Fig. 5. Left and right direction

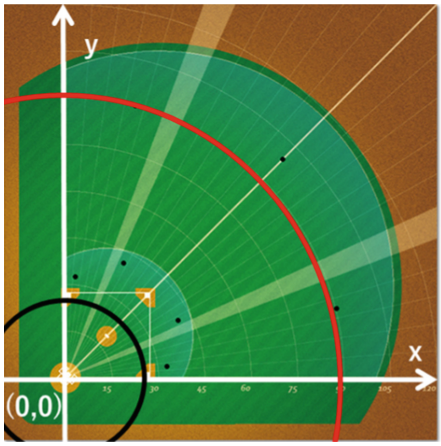

Fig. 6. Short or long distance (Color figure online)

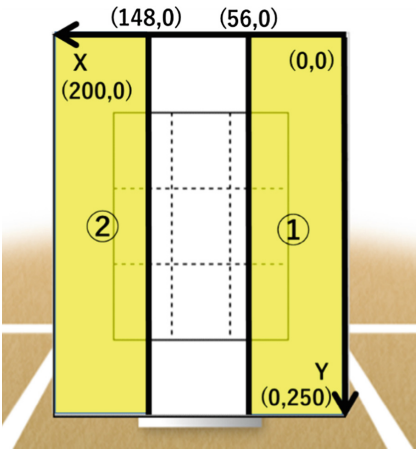

Fig. 7. Inside or outside course

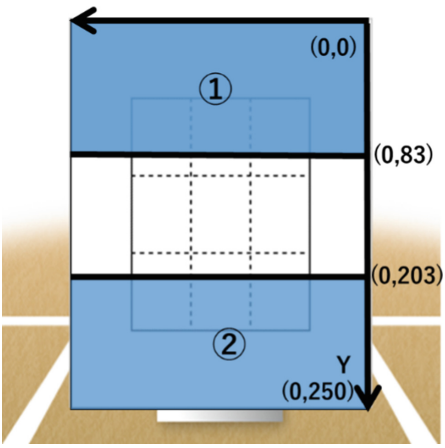

Fig. 8. Higher or lower course

and 0 otherwise. We redefine Slider and Cutter as bending ball, and Split and Off-speed pitch as falling ball. We redefine speed of the ball as fast-ball or slow-ball. Divide the ball speed into three so that the pitching proportion is equal. Among them, we name fast or slow ones as fast-ball and slow-ball. Fast-ball classifies more than 143 km/h. Slow-ball classifies less than 132 km/h. We redefine batting result classification to hits and outs. Variable out is set to 1 when batting result classification is 5. In the same way, variable hit is set to 1 when batting result classification is 4. This is the end of organizing the data.

3 Factor Analysis

In this study, factor analysis and covariance structure analysis are performed. In the factor analysis, we extracted factor of pitcher and factor of batter. Usage variables are summarized in the Table 3 below.

Table 3. Usage variables

Usage variables (batter)
Swing, fly, grounder, long distance, short distance, left direction, right direction
Usage variable (pitcher)
One-bound ball, lower, higher, inside, outside, falling ball, bending ball, shinker, fast-ball, slow - ball

Maximum likelihood method was used for factor analysis method. Criteria for extracting factors were adopted by Gutman-Kaiser criteria. The Gutman-Kaiser criterion is a concept that adopts only factors whose eigenvalues are 1 or higher. Therefore, the yellow part of the table is adopted Considering the accumulation of the variance

from Tables 4 and 5, it is understood that two factors of the batter explain the whole 69% of the total and four factors of the pitcher explain the whole 63% of the total.

Table 4. Extraction result (batter)

Factor	Sum	Variance (%)	Accumulation (%)
1	3.201	45.725	45.725
2	1.664	23.766	69.491
3	0.918	13.109	82.6
4	0.545	7.78	90.38
5	0.375	5.355	95.735
6	0.229	3.277	99.012
7	0.069	0.988	100

Table 5. Extraction result (pitcher)

Factor	Sum	Variance (%)	Accumulation (%)
1	2.387	23.872	23.872
2	1.571	15.706	39.578
3	1.357	13.574	53.152
4	1.076	10.763	63.915
5	0.837	8.368	72.283
6	0.694	6.936	79.219
7	0.643	6.434	85.653
8	0.585	5.851	91.504
9	0.508	5.08	96.584
10	0.342	3.415	100

In the batter's factor, two factors could be extracted as a result of using seven variables. In the pitcher's factor, four factors could be extracted as a result of using ten variables. Each factor matrix is shown in the following Tables 6 and 7.

Table 6. Factor matrix (batter)

Factor matrix	Factor 1	Factor 2
Fly tendency	0.928	−0.141
Short distance	−0.809	−0.549
Swing	0.638	0.462
Right direction	0.572	
Long distance	0.483	
Grounder tendency		0.97
Left direction	0.148	0.571

Table 7. Factor matrix (pitcher)

Factor matrix	Factor 1	Factor 2	Factor 3	Factor 4
Fast-ball	−0.932	−0.137	−0.152	
Slow -ball	0.564			
Bending ball	0.482		−0.267	−0.279
Lower		0.881		
Higher		−0.567		0.1
One-bound ball		0.378	0.145	
Falling ball		0.216	0.975	
Inside		−0.11		0.559
Outside		0.248	−0.144	−0.464
Shinker				0.452

We named factor 1 on the Table 6 as upper swing, factor 2 on the Table 6 as down swing, factor 1 on the Table 7 as a ball to escape outside, factor 2 on the Table 7 as a throw ball to a lower, factor 3 on the Table 7 as a falling ball, factor 4 on the Table 7 as an attack inside. Upper swing chose the name mainly from the fly tendency and right direction. Since the ball is struck from the bottom, the tip of the bat is liable to be delayed. In other words, upper swing is that because the bat is detouring, the point of swing is likely to be delayed. Down swing chose the name mainly from the grounder tendency and left direction. It is the swing to the point in the shortest way to the point, so the point is hard to delay. Ball to escape outside chose the name mainly from the bending ball and ball-speed.

4 Covariance Structure Analysis

In this analysis, we can see how factors extracted by factor analysis affect different variables. This time, we analyze the relationship between the factors and variable hit and out. Maximum likelihood method was also used for covariance structure analysis. The analysis results are summarized in the following Figs. 9 and 10.

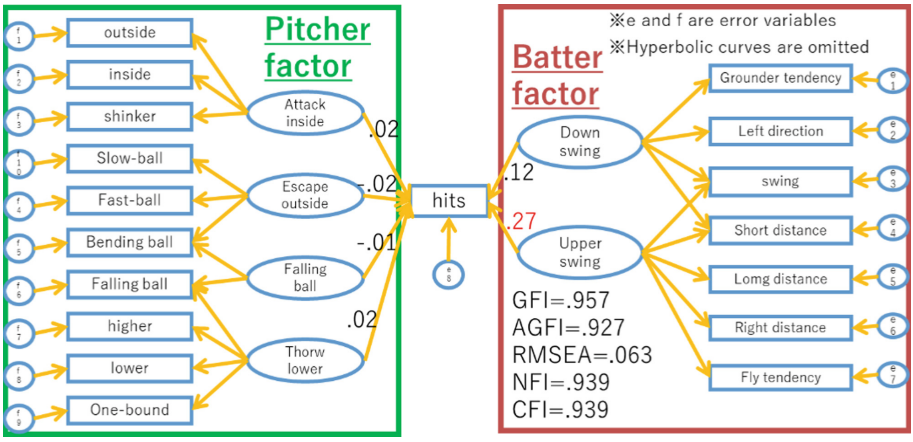

Fig. 9. Factor analysis of hits

An elliptical one is called a latent variable, and a rectangular one is called an observation variable. The latent variable refers to upper swing or down swing, and observation variable refers to left direction, right direction and so on. Latent variables refer to virtual variables such as upper swing or lower swing, and observation variables are variables of actual data such as left direction and right direction. GFI can see whether the total variance of the saturation model can be explained well by the variance of the estimation model. In other words, GFI shows the fit of the model. GFI is a numerical value between 0 and 1, and it is generally said that GFI should be higher than 0.9. NFI can see how much the deviation between the saturation model and the

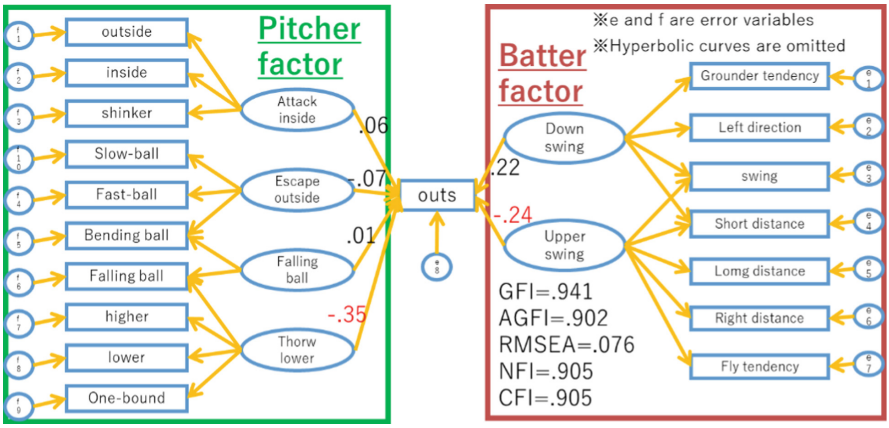

Fig. 10. Factor analysis of outs

estimation model is superior to that of the independent model. NFI is a numerical value between 0 and 1, and it is generally said that NFI should be higher than 0.95. However, GFI and NFI tend to become easier to rise as the model becomes more complicated. This is because the degree of freedom decreases as the model becomes more complicated. Therefore, we use AGFI and CFI in addition to GFI and NFI to judge whether the model is good or bad. AGFI is GFI which takes the degree of freedom into account. AGFI is also a numerical value between 0 and 1, and it is generally said that AGFI should be higher than 0.9. CFI is NFI which takes the degree of freedom into account. CFI is a numerical value between 0 and 1, and it is generally said that CFI should be higher than 0.95. RMSEA is Indicates whether it is divergent between model distribution and true distribution. RMSEA is a numerical value between 0 and 1, and it is generally said that RMSEA should be lower than 0.05. It is said to be a bad model when it exceeds 0.1.

By evaluating Figs. 9 and 10 from the above five observation points, it is understood that values other than RMSEA are satisfied. In other words, you can see that the fit of the model is good.

Arrows indicate the degree of influence between variables in Figs. 9 and 10. The degree of influence is numeric of value between −1 and 1.

It is understood that both the upper swing and the down swing have influenced on hits from Fig. 9. When comparing the two factors, we can see that the influence given by the upper swing a little is great.

It is understood that the upper swing has a negative influence and the lower throw gives a positive influence outs from Fig. 10.

Create latent variable batting rate from observation variables hits and outs. Finally, we analyze the relationship between the factors and this variable. The analysis results are summarized in the following figure.

By evaluating Fig. 11 from the five observation points, it is understood that values other than RMSEA are satisfied. In other words, we can see that the fit of the model is also good.

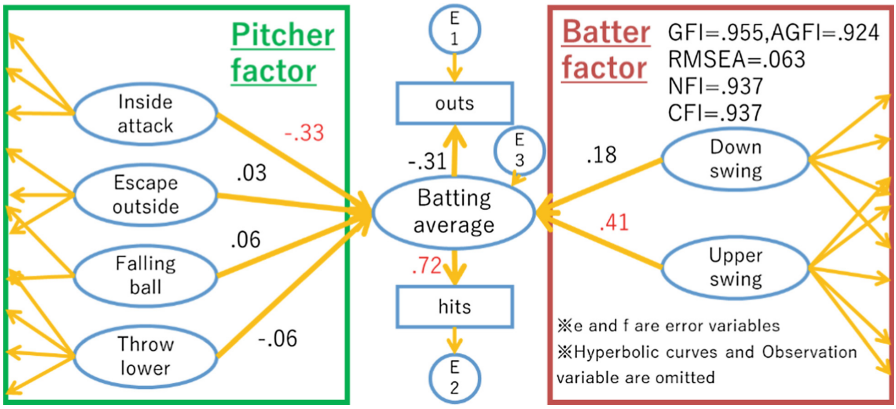

Fig. 11. Factor analysis of batting average

It is understood that the inside attack has a negative influence and the upper swing gives a positive influence outs from Fig. 11.

5 Discussion/Summary

We analyzed the factor analysis of the batting average. The result proved that upper swing is important to raise batting average. We verify the result from the professional baseball season data of 2017. The target is twenty-nine right-handed batters who reached the Stipulated At-bat in Japanese professional baseball. In the case of right pitcher vs. right batter, the batting average was compared with the fly batter and other players. A fly batter refers to a player whose fly rate exceeds the sum of the grounder rate and the liner rate. First, a t-test was performed, but significance was not noticed. Next, we decided to compare using the box plot diagram. The results are summarized in Fig. 12 below.

Box whiskers can see the maximum value, the minimum value, the proportion of the interquartile range. The red line is average of batting average. We can see no difference of average of batting average among fly batter and others. From this figure.

It can be seen that the range from the maximum value to the second quartile of the batting the average of fly batter is narrow. In other words, the fly batter shows that there are many batters with high batting rate compared with other batters.

The future task can be to analyze also combinations other than right-handed pitcher versus right-handed batter who could not be done this time. In addition, we clarify the explanatory variable which has the most influence on improving batting average among

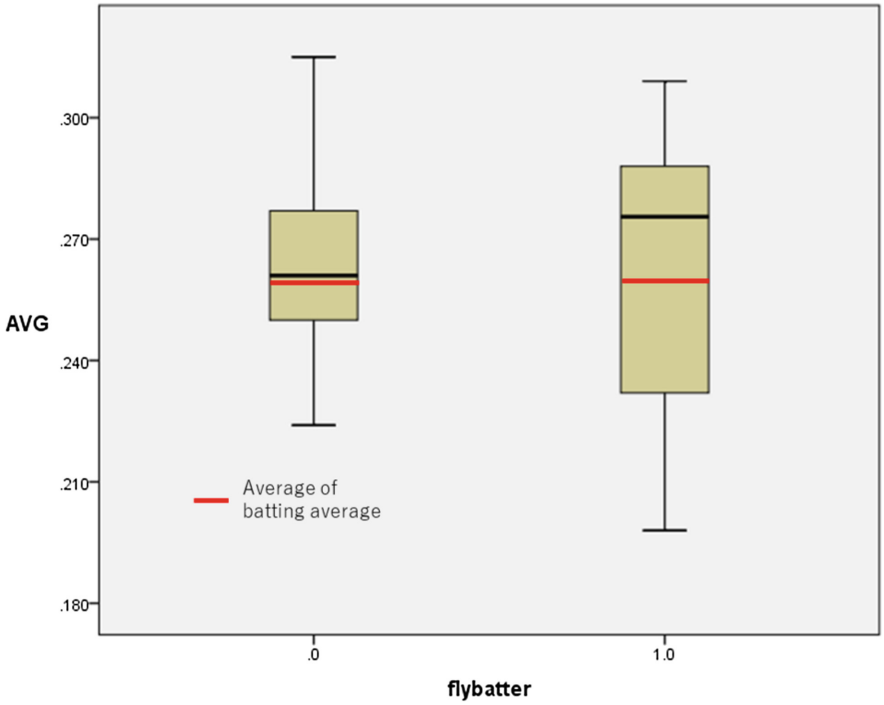

Fig. 12. Comparing batting average fly batter and others (Color figure online)

latent variable upper swing by changing the direction of the arrow of the model in the figure.

References

1. Toyoda, H.: Covariance structure analysis [Amos] -Structural equation modeling (2011)
2. Toyoda, H.: Covariance structure analysis [primer] -Structural equation modeling (2006)
3. Oshio, A.: Psychological and survey data analysis by SPSS and Amos (2005)
4. Bollen, K.A.: Structural Equations with Latent Variables. Wiley, Hoboken (1989)
5. Konno, K.: Structural Equation Modeling - Model Construction Reexamination. (Shizuoka Physics University). http://www.mizumot.com/method/2012-05_Konno.pdf
6. Kano, H.: Fundamentals and actuality of covariance structure analysis—Fundamentals. (Graduate School of Human Studies, Osaka University). http://csrda.iss.u-tokyo.ac.jp/seminar2002_1.pdf. Accessed 19 Feb 2018
7. Moriyasu, Y.: Covariance structure analysis Psychological data analysis exercise. (Graduate School of Engineering). http://cogpsy.educ.kyoto-u.ac.jp/personal/Kusumi/datasem07/moriyasu.pdf. Accessed 19 Feb 2018
8. 1.02 Essence of Baseball. http://1point02.jp/op/index.aspx. Accessed 19 Feb 2018
9. Statistical data of Japan and the world. https://toukeidata.com/index.html. Accessed 19 Feb 2018