

Article

A Study of Winning Percentage in the MLB Using Fuzzy Markov Regression

Seung Hoe Choi ^{1,*} and Seo-Kyung Ji ²¹ School of Liberal Arts and Science, Korea Aerospace University, Goyang 10540, Republic of Korea² Department of Smart Drone Engineering, Korea Aerospace University, Goyang 10540, Republic of Korea; jskyung78@kau.kr

* Correspondence: shchoi@kau.ac.kr

Abstract: In this study, we analyze the winning percentage of 16 teams that have participated in Major League Baseball since 1901. First, 69 variables for each team are classified into pitching, batting, and fielding using factor analysis, and then the effect of the newly classified variables on the winning percentage is analyzed. In addition, after expressing each team's winning rate as a fuzzy number using a fuzzy partition, the linear relationship between the previous year and the next year using the fuzzy probability is investigated, and we estimate the fuzzy regression model and Markov regression model using the Double Least Absolute Deviation (DLAD) method. The proposed fuzzy model describes variables that affect the winning percentage of the next year according to the winning rate of the previous year. The estimated fuzzy regression model showed that the on-base percentage allowed by the pitcher and the on-base percentage of the batter had the greatest effect on the winning percentage.

Keywords: winning percentage; factor analysis; fuzzy partition; DLAD method; fuzzy Markov regression model

MSC: 62A86

Academic Editor: Wanyang Dai

Received: 18 February 2025

Revised: 14 March 2025

Accepted: 17 March 2025

Published: 20 March 2025

Citation: Choi, S.H.; Ji, S.-K. A Study of Winning Percentage in the MLB Using Fuzzy Markov Regression. *Mathematics* **2025**, *13*, 1008. <https://doi.org/10.3390/math13061008>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In ball sports, like soccer, baseball, basketball, and volleyball, which are popular worldwide, one of the key indicators is the winning percentage. This is because the winning rate is directly linked to the revenue of professional teams and the salaries of the players [1,2]. Professional sports teams often strengthen their rosters and reorganize their coaching staff at the end of a season to improve their winning rate. Team restructuring is based on past records of the team or individual players. Hence, finding optimal conditions to enhance the winning percentage based on past performance helps increase the team's revenue [3,4].

Two methods commonly used to study team winning rates in sports are time series analysis and regression analysis. Time series analysis is a method that finds statistical patterns in the winning rate based on past performance [5]. Regression analysis, on the other hand, is a method for finding statistical models that link explanatory variables, which influence the winning rate, to the response variable, which is the winning rate itself [6]. Selecting explanatory variables that affect the response variable is crucial in predicting the winning rate using regression analysis.

Professional baseball is a popular sport in the United States, Japan, and South Korea. The American Major League Baseball (MLB), which began in 1886, currently has 30 teams.

The Japanese and Korean professional baseball leagues, which started in 1934 and 1982, respectively, have 12 teams and 10 teams each [7–9]. Valero has conducted studies on the winning percentage of American professional baseball teams, K. Osawa and K. Aida on Japanese professional baseball teams, and Lee and Kim on Korean professional baseball teams [10–12].

This study examines the winning percentage of 16 teams: 15 teams that participated in MLB since 1901, and the New York Yankees, which joined MLB in 1903. Due to the impact of COVID-19, this study uses data provided by the MLB office and shared by Fangraphs until 2019 [13,14]. Using factor analysis, 68 variables for each team were classified into pitching, batting, and defense, and then the impact of these newly classified variables on the winning rate was analyzed. To achieve this, a fuzzy partition was used to express the winning rate as a fuzzy number, and the relationship between the winning rate of the previous and following years was examined using fuzzy probability. The Markov regression model was employed to identify variables that influence the winning rate of the following year based on the previous year's winning rate.

This study explains the regression model for the win percentage, which was limited by mathematical rigor, together with batting, pitching, and fielding power. Pitching, hitting, and fielding power cannot be said to be independent of each other in predicting the winning rate of professional baseball, and variables expressed as numbers and ratios can be dependent on each other [10–12]. Under these conditions, it is necessary to consider a fuzzy regression model for the winning rate without separating the number and ratio of attack, pitching, and defense power. In addition, investigating and confirming how the progress of the team maintained through the records of last season affects the new season is a necessary process to improve the team's winning rate. So, we propose a fuzzy Markov regression model for the future win rate based on the past win rate using fuzzy numbers.

This study is structured as follows. Section 2 introduces the factor analysis for the cluster of variables, the fuzzy partition, and the double least absolute deviation method to estimate the fuzzy regression model for the winning rate. Section 3 proposes the fuzzy regression and fuzzy Markov regression model and explains the variables that influence the change in the following year's winning rate based on the previous year's winning rate. It also identifies the variables that have the most significant impact when the winning rate increases or decreases compared to the past. The conclusion of this study is presented in the final section.

2. Methodology

2.1. Factor Analysis

The 68 variables selected to explain the winning rates are shown in Table 1. The abbreviations used in Table 1 are based on the data provided by the MLB office [15]. Among the explanatory variables listed in Table 1, some are recorded as counts and others as ratios. These variables were categorized into batting, pitching, and fielding categories, and further divided into count variables (vac) and ratio variables (var). Factor analysis was performed to group variables with high correlation within each category [16]. Using the statistical software SPSS 28, the factor analysis results showed a Kaiser–Meyer–Olkin measure of 0.73 and a p -value of less than 0.0001 for Bartlett's test of sphericity. To determine the number of groups, an eigenvalue threshold of 0.8 was set, and factor analysis was conducted using the principal component analysis and varimax methods [16]. The results of the factor analysis are presented in Table 1. Eleven variables ($b_c^1, \dots, b_c^4, p_c^1, \dots, p_c^4, f_c^1, \dots, f_c^3$) were categorized as count variables, while seven variables ($b_r^1, \dots, b_r^3, p_r^1, \dots, p_r^3, f_r^1$) were categorized as ratio variables. In Table 1, OBP and WHIP indicate the Online Browsing

Platform and Walks plus Hits divided by Innings Pitched, respectively [14,15]. The meaning of variables introduced in Table 1 can be found in MBL and Fangraphs homepage [7,14].

Table 1. Results of factor analysis and explanatory variables.

32 Batting Variables	11 vac	b_c^1	RBI, BB, SB, E	-Variables excluded as a result of discriminant analysis: G, PA, AB, R, CS, OPS, Fld%, HBP, SF, SH, Def, IBB
		b_c^2	SO, DP	
		b_c^3	HR, 2B	
		b_c^4	3B, H, Sh	
	9 var	b_r^1	BA, OBP, wOBA, wRC+	
		b_r^2	ISO, SLG	
		b_r^3	OFF, WAR, BsR	
29 Pitching Variables	10 vac	p_c^1	BK, HBP, BB, WP	-Variables excluded as a result of discriminant analysis: G, IP, R, ER, SO, HR9, DP, Fld%, Page, ShO, GS, IBB, E
		p_c^2	CG, SV, tSHO	
		p_c^3	H, HR	
		p_c^4	SO9	
	6 var	p_r^1	ERA, FIP, WHIP	
		p_r^2	BABIP, LOB%	
		p_r^3	WAR	
8 Fielding Variables	3 vac	f_c^1	E	-Variables excluded as a result of discriminant analysis: PO, PB, SB, CS
		f_c^2	DP	
		f_c^3	A	
	1 var	f_r^1	DEP	

2.2. Fuzzy Partition on the Winning Percentage

To investigate the explanatory variables affecting the winning percentages of 16 MLB teams, this study standardized the winning percentages $\left(w_p^k(i)\right)$ of each team. Table 2 presents the descriptive statistics of the winning percentages.

Table 2. Descriptive statistics for the winning percentage.

N	Minimum	Mean ($\bar{x}$)	Median	Maximum	Variance	Standard Deviation(s)
1902	0.235	0.506	0.512	0.763	0.007	0.085

Table 2 shows the central tendency and dispersion of 1902 (N) data on the winning percentage. In addition, standardization of the data presented in Table 1 is necessary to efficiently compare the effects of variables on the winning rate. The standardized winning rate is given by the following:

$$S_{w_p}^k(i) = \frac{w_p^k(i) - \bar{x}}{s}, \quad k = 1, \dots, 16, \quad i = 1, \dots, 119 \quad (1)$$

In Equation (1), k and i denote the team and the year, respectively. In this study, the explanatory variables affecting the winning percentage were also standardized for analysis.

To probabilistically express the influence of the previous year's winning rate on the next year's winning rate, it is necessary to convert the data into categorical forms. Fuzzy partitioning can be used for categorizing the winning percentage [17].

Fuzzy partitioning involves decomposing the entire set using fuzzy numbers [18]. A fuzzy set is essential for mathematically defining sets with unclear boundaries, such as high winning rates or low offensive power.

The fuzzy set, $A = \{(x, \mu_A(x)) : x \in X\}$, first proposed by Zadeh (1965), is defined as follows: the membership function of set A is $\mu_A : X \rightarrow [0, 1]$. For any element α in the interval $[0, 1]$, the α -level set is $A_\alpha = \{x \in X : \mu_A(x) \geq \alpha\}$ [18]. The membership function of a triangular fuzzy number $A = (l_a, a, r_a)_T$ is defined as follows:

$$\mu_A(x) = \begin{cases} 1 - \left(\frac{a-x}{l_a}\right) & \text{for } 0 \leq a-x \leq l_a \\ 1 - \left(\frac{x-a}{r_a}\right) & \text{for } 0 \leq x-a \leq r_a \\ 0 & \text{for otherwise} \end{cases} \quad (2)$$

In Equation (2), a is referred to as the mode, and l_a and r_a are the left and right end points of the fuzzy number A, respectively. Specifically, if the membership function satisfies $L(a-x) = R(a+x)$, it is called a symmetric fuzzy number and is denoted as $(a, s)_T$ and is $s = l_a = r_a$ [18–20].

If the set $\{A_j : j = 1, \dots, k\}$ is a fuzzy partition of the set $[m, M]$, it must satisfy the following condition.

$$\begin{aligned} (1) & m = a_1, M = a_k, A_i(a_i) = 1 \\ (2) & A_1 = (a_1, a_1, a_2)_T, A_i = (a_{i-1}, a_i, a_{i+1})_T \ (i = 2, \dots, k-1), A_k = (a_{k-1}, a_k, a_k)_T \\ (3) & A_i(x) = 0 \text{ if } x \notin (a_{i-1}, a_{i+1}), A_i(x) \nearrow \text{ on } [a_{i-1}, a_i], A_i(x) \searrow \text{ on } [a_i, a_{i+1}] \end{aligned} \quad (3)$$

In Equation (3), any element x in the interval $[m, M]$ is given by the following:

$$\sum_{i=1}^{n+1} A_i(x) = 1 \quad (4)$$

This satisfies the condition that the sum of the membership function values for any point x in the set X equals 1, as expressed in Equation (4) [17,18].

The maximum and minimum values of the standardized winning percentages $S_{wp}(i)$ are -3.19 and 3.03 , respectively. Therefore, the fuzzy partition for the interval $[-3.19, 3.03]$ is as follows.

$$\begin{aligned} l_1 &= 2.508, i_2 = 0.757, l_3 = 2.654, l_4 = 0.3 \\ a_1 &= -3.19, a_i = a_{i-1} + l_i \ (2 \leq i \leq 4), a_5 = 3.03 \end{aligned} \quad (5)$$

Figure 1 shows the fuzzy partitioning using the center a_i and the width l_i given in Equation (5).

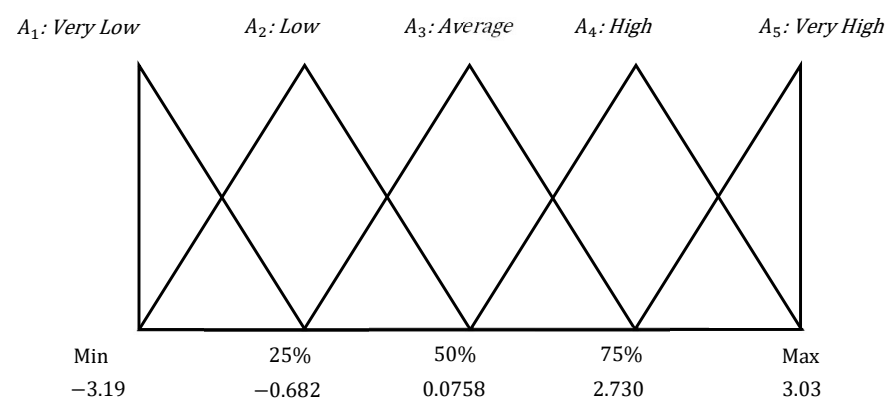


Figure 1. Fuzzy partition for standardized winning percentage.

This study investigates the change from the fuzzy number $A_i^{(b)}$, representing the winning rate of the previous year, to the fuzzy number $A_i^{(c)}$, representing the winning rate of the following year.

2.3. Fuzzy Probability of Future Win Percentage

It is meaningful to study the future winning rate that changes according to the past winning rate. The reason is that the future can be predicted according to today's winning rate. It is confirmed that the past and present winning rates are dependent, and the probability of the future winning percentage according to the past winning rate expressed as a fuzzy number is calculated. The future win rate change according to the past win rate is expressed as a fuzzy probability.

The process by which the win rate of the previous year changes to the win rate of the following year can be observed in Table 3. Since win rates prior to 1901 are omitted in Table 3, the total sum is 1886. Additionally, Table 3 shows that the frequency of the cells located on the diagonal (x_{ij}) and adjacent cells (x_{i-1i} or x_{i+1i}) is higher than that of other cells (x_{ij} , $|i - j| > 1$).

Table 3. Contingency table.

Before	$A_1^{(c)}$	After				Total
		$A_2^{(c)}$	$A_3^{(c)}$	$A_4^{(c)}$	$A_5^{(c)}$	
$A_1^{(b)}$	17	44	4	2	0	67
$A_2^{(b)}$	39	345	149	91	0	624
$A_3^{(b)}$	9	147	165	162	5	488
$A_4^{(b)}$	0	89	167	391	24	671
$A_5^{(b)}$	0	1	2	26	7	36
Total	65	626	487	672	36	1886

The likelihood ratio test for Table 3 explains the dependence of the winning rate on the past and the future ($X^2(16) = 636$, $p = \text{value} < 0.001$). Gamma statistics ($\lambda = 0.486$, $p\text{-value} < 0.001$) showed a significant and positive relation between the two variables [16].

In Table 3, we define $P_{ij} = \frac{x_{ij}}{x_{i.}}$, $x_{i.} = \sum_{j=1}^5 x_{ij}$, and $s_{ij}(\alpha) = z_{\frac{\alpha}{2}} \sqrt{\frac{p_{ij}(1-p_{ij})}{x_{i.}}}$. Here, $z_{\frac{\alpha}{2}}$ means the $(1 - \alpha)100\%$ of the area under a normal curve. Then the left and right ends of the $(1 - \alpha)100\%$ confidence interval for p_{ij} are $l_{p_{ij}}(\alpha) = p_{ij} - s_{ij}(\alpha)$ and $r_{p_{ij}}(\alpha) = p_{ij} + s_{ij}(\alpha)$, respectively.

The α -level set of the element P_{ij} in the fuzzy transition probability matrix, Table 4, for past and future win percentage is $P_{ij}(\alpha) = [l_{p_{ij}}(\alpha), r_{p_{ij}}(\alpha)]$. Additionally, $l_{p_{ij}}(0) = \max\{0, p_{ij} - \alpha \cdot s_{ij}(\alpha)\}$ and $r_{p_{ij}}(0) = \min\{p_{ij} + \alpha \cdot s_{ij}(\alpha), 1\}$ give the fuzzy probability $P_{ij} = (l_{p_{ij}}(0), p_{ij}, r_{p_{ij}}(0))_T$ [19]. Table 4 shows that the past winning rate does not change significantly from the past winning rate except when the past winning rate is very low or very high.

Table 4. Fuzzy transition probability matrix.

	$A_1^{(c)}$	$A_2^{(c)}$	$A_3^{(c)}$	$A_4^{(c)}$	$A_5^{(c)}$
$A_1^{(b)}$	$(0.15, 0.26, 0.37)_T$	$(0.48, 0.60, 0.72)_T$	$(0.05, 0.14, 0.22)_T$	$(0.00, 0.00, 0.00)_T$	$(0.00, 0.00, 0.00)_T$
$A_2^{(b)}$	$(0.05, 0.07, 0.09)_T$	$(0.51, 0.55, 0.59)_T$	$(0.20, 0.23, 0.27)_T$	$(0.11, 0.14, 0.17)_T$	$(0.00, 0.00, 0.00)_T$
$A_3^{(b)}$	$(0.00, 0.01, 0.02)_T$	$(0.27, 0.31, 0.35)_T$	$(0.30, 0.34, 0.38)_T$	$(0.30, 0.34, 0.39)_T$	$(0.00, 0.00, 0.01)_T$
$A_4^{(b)}$	$(0.00, 0.00, 0.01)_T$	$(0.11, 0.14, 0.16)_T$	$(0.21, 0.24, 0.27)_T$	$(0.54, 0.58, 0.62)_T$	$(0.02, 0.04, 0.05)_T$
$A_5^{(b)}$	$(0.00, 0.00, 0.00)_T$	$(0.00, 0.00, 0.00)_T$	$(0.03, 0.14, 0.25)_T$	$(0.51, 0.67, 0.82)_T$	$(0.07, 0.19, 0.32)_T$

2.4. Fuzzy Markov Regression

Regression analysis is a statistical method used to estimate a regression function that explanation or prediction the relationship between explanatory variables, classified through factor analysis, and win rates expressed in fuzzy numbers [16]. In this study, the influence of explanatory variables on the process by which the win rate status of the previous year changes to the status of the next year is examined using the estimated regression model. In this section, a fuzzy regression model for a standardized winning rate (S_{wp}) is estimated. The effect on the winning rate is investigated using a fuzzy regression model for the number part, the ratio part, and the whole. In addition, the effect on the winning rate of the following year, which is conditional on the winning rate of the previous year, is explained through an estimated fuzzy regression model.

In this study, win rates are expressed as fuzzy numbers through fuzzy partitioning. Therefore, a fuzzy regression model is used to estimate the relationships between these fuzzy numbers [20–23]. The fuzzy regression model for the triangular fuzzy number set $\{(Y_i, X_i, \dots, X_{ip}) : i = 1, \dots, n\}$ is given by the following:

$$Y_i = A_0 + A_1 X_{i1} + \dots + A_{ip} X_{ip} + E_i \quad (6)$$

In Equation (6),

$$Y_i = (l_e(y_i), y_i, r_e(y_i))_T, \\ A_i = (l_e(a_i), a_i, r_e(a_i))_T, \quad X_{ij} = (l_e(x_{ij}), x_{ij}, r_e(x_{ij}))_T$$

and $E_i = (l_e, e_i, r_e)_T$ is fuzzy error. Choi and Buckley (2008) employed the following three-step method to estimate the fuzzy regression model (6) [20].

Step 1. Calculate center of 0-level set

$$m_y = \frac{y_i + l_e(y_i) + r_e(y_i)}{3}, \quad m_{x_{ij}} = \frac{x_{ij} + l_e(x_{ij}) + r_e(x_{ij})}{3} \quad (7)$$

Equation (7) calculates the centroids of the response variable and the explanatory variables.

Step 2. Estimate the regression model of center $\{(m_y, m_{x_{i1}}, \dots, m_{x_{ip}}) : i = 1, \dots, n\}$

$$\sum_{i=1}^n \left| m_y - \sum_{k=0}^p a_k m_{x_{ij}} \right| + \sum_{k=0}^p |a_k| \equiv \text{Min!}, \text{ denoted by } \hat{a}_k \quad (8)$$

In the Equation (8), $\hat{a}_k (k = 0, \dots, p)$ is referred to as the DLAD (Doble Least Absolute Deviation) method.

Step 3. Estimate fuzzy error $E_i = (e_i, l_\epsilon, r_\epsilon)_T$

$$\begin{aligned} \sum_{i=1}^n \left| l_\epsilon(y_i) - \sum_{k \in P_0} \hat{a}_k l_\epsilon(x_{ik}) - \sum_{k \in N_\epsilon} \hat{a}_k r_\epsilon(x_{ik}) - l_\epsilon \right| &\equiv \text{Min!}, \text{ denoted by } \hat{l}_\epsilon \\ \sum_{i=1}^n \left| r_\epsilon(y_i) - \sum_{k \in P_0} \hat{a}_k l_\epsilon(x_{ik}) - \sum_{k \in N_\epsilon} \hat{a}_k l_\epsilon(x_{ik}) - r_\epsilon \right| &\equiv \text{Min!}, \text{ denoted by } \hat{r}_\epsilon \end{aligned} \quad (9)$$

When the fuzzy numbers of the response variable are symmetric, the estimated error from Equation (9) satisfies $\hat{l}_\epsilon = \hat{r}_\epsilon$.

The DLAD method regulates the regression coefficient in the least absolute deviation method, which estimates using the absolute value of the error. This reduces the variance in the model and finds variables with small influence [20].

3. Results and Discussion

3.1. Fuzzy Regression for Winning Percentage

The fuzzy regression model estimated using the DLAD method (8,9) and the data classified by Table 1 of the 16 teams participating since 1901 is shown in (10).

$$\begin{aligned} S_{wp}(i) = & 0.002 + 0.075b_c^1(i) - 0.043b_c^2(i) + 0.1b_c^3(i) + 0.041b_c^4(i) \\ & - 0.006p_c^1(i) + 0.172p_c^2(i) - 0.0081p_c^3(i) - 0.055p_c^4(i) - 0.106f_c^1(i) \\ & + 0.111f_c^2(i) - 0.051f_c^3(i) - 0.369b_r^1(i) + 0.318b_r^2(i) - 0.024b_r^3(i) \\ & - 0.546p_r^1(i) - 0.274p_r^2(i) + 0.018p_r^3(i) + 0.04f_r^1(i) + (0, 0.306)_T \end{aligned} \quad (10)$$

The estimated fuzzy regression model (10), estimated by the DLAD method, is that the batting part b_r^1 (BA, OBP, wOBA, wRC+) and b_r^2 (ISO, SLG), expressed as ratios, have the greatest positive effect on the winning percentage, and the pitching power part p_r^1 (ERA, FIP, WHIP) and p_r^2 (BABIP, LOB%), expressed as ratios, have the greatest negative effect on the winning rate. This result shows that the proposed fuzzy regression model (10) contains the results of the regression equation for team score and ERA [24,25]. The proposed fuzzy regression model (10) shows that the problem that may arise due to the dependence of variables used in baseball games has been overcome.

Unlike the estimated regression model (10), the fuzzy regression model was estimated by separating the variable represented by the number (vac) and the variable represented by the ratio (var). The fuzzy regression models for the win rate estimated independently using the DLAD method (8,9) are shown in (11) and (12), respectively.

$$\begin{aligned} S_{wp}^c(i) = & 0.013 + 0.124b_c^1(i) - 0.137b_c^2(i) + 0.331b_c^3(i) + 0.103b_c^4(i) \\ & - 0.086p_c^1(i) + 0.298p_c^2(i) - 0.291p_c^3(i) + 0.214p_c^4(i) \\ & - 0.232f_c^1(i) + 0.262f_c^2(i) - 0.042f_c^3(i) + (0, 0.3)_T \end{aligned} \quad (11)$$

The estimated regression model (11) shows that, among the variables based on the counting, p_c^2 (CG, SV, tSHO) and p_c^3 (H, HR) have the greatest influence on the winning rate. In addition, the regression model (12) shows that p_r^1 and b_r^1 have the greatest influence on the winning percentage among variables expressed as ratios.

$$\begin{aligned} S_{wp}^r(i) = & -0.005 + 0.599b_r^1(i) + 0.38b_r^2(i) + 0.024b_r^3(i) \\ & - 0.598p_r^1(i) - 0.364p_r^2(i) + 0.069p_r^3(i) + 0.09f_c^1(i) + (0, 0.295)_T \end{aligned} \quad (12)$$

3.2. Fuzzy Markov Regression for Winning Percentage

Based on contingency in Table 3, it is possible to estimate 25 Fuzzy Markov regression models [26]. In this study, regression models will be estimated for the two states with the highest proportions and fuzzy probability in Tables 3 and 4: $(A_5^{(b)} \rightarrow A_4^{(c)}, 72.2\%$,

$(0.51, 0.67, 0.82)_T$) and $(A_1^{(b)} \rightarrow A_2^{(c)}, 65.7\%, (0.48, 0.60, 0.72)_T)$. In the case of a change from the previous year $A_5^{(b)}$ to the next year $A_4^{(c)}$, the Fuzzy Markov regression model estimated using the DLAD method (8,9) is as follows.

$$S_{wp}^{c(5,4)}(i) = 0.563 + 0.04b_c^1(i) + 0.168b_c^2(i) + 0.2321b_c^3(i) + 0.086b_c^4(i) - 0.171p_c^1(i) + 0.183p_c^2(i) - 0.237p_c^3(i) + 0.27p_c^4(i) + 0.177f_c^1(i) - 0.166f_c^2(i) - 0.237f_c^3(i) + (0, 0.55)_T \quad (13)$$

Equation (13) for the vac demonstrates the influence of explanatory variables based on the number of transitions when the win rate changes from the “very high” stage to the “high” stage. Equation (13) demonstrates that, when a pitcher allows a hit or a home run $p_c^3(i)$, or when a base runner is put out $f_c^3(i)$, the win rate decreases compared to the previous year.

By similar method to Equation (11), we obtained the following fuzzy Markov regression model for the var from $A_5^{(b)}$ to $A_4^{(c)}$.

$$S_{wp}^{r(5,4)}(i) = 0.55 + 0.641b_r^1(i) + 0.2b_r^2(i) - 0.172b_r^3(i) - 0.496p_r^1(i) - 0.151p_r^2(i) + 0.041p_r^3(i) + 0.143f_r^1(i) + (0.0.642)_T \quad (14)$$

In Equation (14), which explains the change in win rate based on proportions, it can be observed that, if the pitcher’s earned run average or runs allowed $p_r^1(i)$ is high, the win rate decreases. Fuzzy Markov Regression Model for vac and var from $A_1^{(b)}$ to $A_2^{(c)}$ is as follows:

$$S_{wp}^{c(1,2)}(i) = -0.63 + 0.07b_c^1(i) - 0.164b_c^2(i) + 0.17b_c^3(i) - 0.067b_c^4(i) + 0.008p_c^1(i) + 0.355p_c^2(i) - 0.055p_c^3(i) - 0.169p_c^4(i) - 0.142f_c^1(i) + 0.103f_c^2(i) - 0.188f_c^3(i) + (0, 0.57)_T \quad (15)$$

Equation (15) explains that the pitcher’s complete games or saves, $p_c^2(i)$, and the batting average or weighted on-base percentage, $b_r^1(i)$, increase the win rate. The results of Equations (13)–(15) share similarities with findings from other research papers [27,28]. The fuzzy Markov regression model (13)–(15), estimated by the DLAD method, shows what each team should do first in the stove league, which is the time for teams in the Major League to prepare for the new season.

Figure 2 below describes the two fuzzy regression models used in this study. In Figure 2, the upward direction refers to the fuzzy regression model of the winning rate for all variables, and the downward direction refers to the fuzzy Markov regression model for the winning percentage that changes from the $A_s^{(b)}$ stage in the past to the $A_t^{(c)}$ stage in the present.

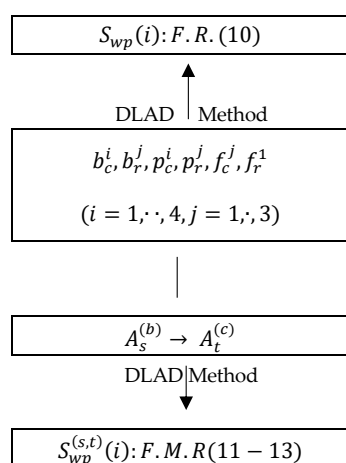


Figure 2. Block diagram for regression model of winning percentage.

Finally, in Figure 2, the upward direction refers to the fuzzy regression model of the winning rate for all variables, and the downward direction refers to the fuzzy Markov regression model for the winning percentage that changes from the $A_s^{(b)}$ stage in the past to the $A_t^{(c)}$ stage in the present.

4. Conclusions

In this paper we investigated the variables that affect the winning percentage by using data of MLB from 1901 to 2019, which is the year before the outbreak of COVID-19. To this purpose, we classified 68 variables into 18 categorical variables using factor analysis, and independently classified data based on number and ratio.

In this study, a fuzzy partition was used to categorize the winning percentage of teams belonging to the MLB. The effect of variables such as batting, pitching, and fielding on the categorized winning percentage change was confirmed using a fuzzy Markov regression model. In particular, the double least absolute deviation method was used to estimate the fuzzy Markov regression model.

In the case of a team whose winning percentage increased, the number of hits, home runs, baserunner kills, and pitcher ERA had the greatest influence on the winning rate. On the other hand, in the case of a team whose winning rate has decreased, the number of shutouts or complete games by pitchers and the batting average by batters had the greatest influence on the winning rate.

This study, which examined the average of 16 teams belonging to the MLB, needs to further study the characteristics of each team. In future research, we will study which variables have the greatest influence on the winning rate of 32 teams in the MLB. We will also study the change in winning percentage according to individual player's ability.

Author Contributions: Conceptualization, S.H.C.; methodology, S.H.C.; software, S.H.C. and S.-K.J.; investigation, S.H.C.; data curation, S.-K.J.; writing—original draft preparation, S.H.C. and S.-K.J.; funding acquisition, S.H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2021R1F1A1057507).

Data Availability Statement: The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mullin, B.J.; Hardy, S.; Sutton, A. *Sport Management; Human Kinetics*: Champaign, IL, USA, 2007.
2. Stimel, D.S. Dependence Relationships between On Field Performance, Wins, and Payroll in Major League Baseball. *J. Quant. Anal. Sports* **2011**, *7*, 1–19.
3. Fort, R. *Sports Economics*, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2010.
4. Kim, M.-C. A Study on the Winning and Losing Factors of Para Ice Hockey Using Data Mining-Based Decision Tree Analysis. *Appl. Sci.* **2023**, *13*, 1334. [CrossRef]
5. Lee, W.; Jhang, H.; Lee, S.; Choi, S.H. Forecasting Winning Rates in Major League Baseball Based on Fuzzy Logic. *J. Korean Inst. Intell. Syst.* **2020**, *30*, 366–372.
6. Stick, J.D. A Regression Analysis of Predictors on the Productivity Indices of Major League Baseball. ProQuest Dissertations and Thesis, The University of Nebraska, Lincoln, NE, USA, 2005.
7. MLB Homepage. Available online: <https://www.mlb.com/> (accessed on 15 January 2023).
8. NPB Homepage. Available online: <https://npb.jp/> (accessed on 20 March 2023).
9. KBL Homepage. Available online: <https://www.koreabaseball.com/> (accessed on 12 February 2023).
10. Soto Valero, C. Predicting Win-Loss outcomes in MLB regular season games—A comparative study using data mining methods. *Int. J. Comput. Sci. Sport* **2016**, *15*, 91–112. [CrossRef]

11. Osawa, K.; Aida, K. The technique of calculating the winning percentage in baseball games with runners' advancement statistics. *Jpn. J. Ind. Appl. Math.* **2008**, *18*, 321–346.
12. Lee, J.T.; Kim, Y.T. A study on the estimation of winning percentage in Korean pro-baseball. *J. Korean Data Anal. Soc.* **2006**, *8*, 857–869.
13. MLB Stats Page. Available online: <https://www.mlb.com/stats/> (accessed on 13 January 2023).
14. Fangraphs Homepage. Available online: <https://www.fangraphs.com/> (accessed on 8 January 2023).
15. MLB Glossary Page. Available online: <https://www.mlb.com/glossary/> (accessed on 8 January 2023).
16. Kim, H.K.; Choi, S.H. *Statistical Analysis with Applications*; Kyung Moon Sa: Seoul, Republic of Korea, 2002.
17. Guillaume, S.; Charnomordic, B.; Loisel, P. Fuzzy partitions: A way to integrate expert knowledge into distance calculations. *Inf. Sci.* **2013**, *245*, 76–95. [[CrossRef](#)]
18. Zadeh, L. Fuzzy Sets. *Inf. Control* **1965**, *8*, 338–353. [[CrossRef](#)]
19. Buckley, J.J. *Fuzzy Probability*; Springer: Berlin/Heidelberg, Germany, 2003.
20. Choi, S.H.; Buckley, J.J. Fuzzy regression using least absolute deviation estimators. *Soft Comput.* **2008**, *12*, 257–263. [[CrossRef](#)]
21. Jung, H.Y.; Yoon, J.H.; Choi, S.H. Fuzzy linear regression using rank transform method. *Fuzzy Sets Syst.* **2015**, *274*, 97–108. [[CrossRef](#)]
22. Phaiboon, S.; Phokharatkul, P. Multi-Boundary Empirical Path Loss Model for 433 MHz WSN in Agriculture Areas Using Fuzzy Linear Regression. *Sensors* **2023**, *23*, 3525. [[CrossRef](#)] [[PubMed](#)]
23. Yoon, J.H.; Choi, S.H. A Large Sample Study of Fuzzy Least-Squares Estimation. *Axioms* **2025**, *14*, 181. [[CrossRef](#)]
24. Philip, B.; Paul, D.B.; Bruce, D.W. Predicting Run Production and Run Prevention in Baseball: The Impact of Sabermetrics. *Int. J. Bus.* **2012**, *2*, 67–75.
25. Cho, Y.S.; Cho, Y.J. A study on winning percentage using batter's runs and pitcher's runs in Korean professional baseball league. *J. Korean Data Anal. Soc.* **2005**, *7*, 2303–2312.
26. Yen, A.M.-F.; Chen, T.H.-H. Mixture Multi-state Markov Regression Model. *J. Appl. Stat.* **2007**, *34*, 11–21.
27. Davis, M.C. The interaction between baseball attendance and winning percentage: A VAR analysis. *Int. J. Sport Financ.* **2008**, *3*, 58–73.
28. Lemke, R.J.; Leonard, M.; Tlhokwane, K. Estimating attendance at Major League Baseball games for the 2007 season. *J. Sports Econ.* **2010**, *11*, 316–348.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.