The Korean Journal of Applied Statistics (2010) DOI: 10.5351/KJAS.2010.23.6.1191
23*(6), 1191–1199*

A Statistical Analysis of Professional Baseball Team Data: The Case of the Lotte Giants

Young-Seuk Cho[1] · Jun-Tae Han[2] · ChanKeun Park[3] · Tae-Young Heo[4]

[1]Department of Statistics, Pusan National University
[2]National Health Insurance Policy Research Institute, National Health Insurance Corporation
[3]Department of Data Information, Korea Maritime University
[4]Department of Data Information, Korea Maritime University

(Received April 2010; accepted October 2010)

Abstract

Knowing what factors into a player's ability to affect the outcome of a sports game is crucial. This knowledge helps determine the relative degree of contribution by each team member as well as sets appropriate annual salaries. This study uses statistical analysis to investigate how much the outcome of a professional baseball game is influenced by the records of individual players. We used the Lotte Giants' data on 252 games played between 2007 and 2008 that included environmental data(home or away games and opponents) as well as pitchers' and batters' data. Using a SAS Enterprise Miner, we performed a logistic regression analysis and decision tree analysis on the data. The results obtained through the two analytic methods are compared and discussed.

Keywords: Decision tree analysis, logistic regression analysis, odds ratio, SAS Enterprise Miner.

1. Introduction

The Korean Baseball League(KBL) was founded in 1982. The Lotte Giants(a professional baseball team based in Busan, Korea) is one of the original franchises of the Korea Baseball Organization(KBO) league. The Lotte Giants won the league championship in 1984 and 1992, and were twice the league runner-ups, in 1995 and 1999. Between 1982 and 2008, they played a total of 3,263 games, 1,464 of which they won, 1,715 of which they lost, and 84 of which they tied, for a winning percentage of 0.449. In 2009, the Lotte Giants defended their title as Korea's most popular professional baseball club and their total attendance increased to 1,380,018.

There is existing literature of statistical analysis on Korean professional baseball data, for instance, Cho and Cho (2003) analyzed the Beane Count of baseball teams and its correlation with the winning average. They performed a cluster analysis and a regression analysis using the data of homeruns scored and allowed, and walks earned and allowed, which are the basic components of the Beane Count. Kim (2004), meanwhile, pointed out how the criteria used for determining team rankings

[4]Corresponding author: Assistant Professor, Department of Data Information, Korea Maritime University, 1 Dongsam-Dong, Yeongdo-Gu, Pusan 606-791, Korea. E-mail: heoty@hhu.ac.kr

in Korean professional baseball and football are not statistically reasonable, giving suggestions for correcting this ranking irrationality.

As for Lee and Kim (2006a), they estimated the winning percentages of Korean professional baseball teams by finding the optimal exponent for the Pythagorean theorem, under the criterion of a root mean squared error. Lee and Kim (2006b) used the same method to estimate the winning percentages of professional women's basketball teams and professional football teams. Lee and Kim (2007) also proposed a model for predicting the winning percentages of professional baseball teams, based on overall team power, rather than points scored and allowed. Cho *et al.* (2007) analyzed 6,146 games, looking at factors such as the field environment(home or away games), offense events(hits, stolen bases, sacrifice bunts, sacrifice fly, warks, hit by pitch, strikeouts, and double plays), defense events(errors), the role of the pitcher(being hit, allowing stolen bases, allowing walks, hitting the batter, and strikeouts), in order to quantify how these variables affect the outcome of a game. More recently, Shin *et al.* (2007) studied the data from the Samsung Lions to investigate the extent to which the outcome of a professional baseball game is influenced by the records of individual players, utilizing statistical analysis techniques and software applications. Lee and Cho (2009) assessed the influence of the home-field advantage in Korean professional baseball using a logistic regression model. A few key references on baseball and prediction are Lindsey (1963), Albright (1993), James *et al.* (1993), Albert (1994), Barry and Hartigan (1994), Choi and Shim (1995), Bennett (1998), and Hong and Choi (2008).

The software application used in this study for statistical analysis is the SAS Enterprise Miner as a data mining tool. Data mining is the process of extracting meaningful knowledge from large sets of data (Han and Kamber, 2001). For a prediction of the odds of winning, we used a logistics regression model and decision tree model. The data used in this study include the 249 games of 252 total games played by the Lotte Giants between 2007 and 2008 which excludes the three tie games. We investigated which of the factors related to batters and pitchers influenced the outcomes of these games, and examined the records of individual batters and pitchers to assess their influence on the outcome of a game.

The rest of this study is organized as follows: In Section 2, we present the results of our statistical analysis of the overall game data of the Lotte Giants. In Section 3, the statistical analysis focuses on the batter data, and in Section 4, on the pitcher data. In the last section, we present the conclusion based on the results of the above analysis.

2. Statistical Analysis of Lotte Giants' Overall Data

The Lotte Giants' game data were obtained from *Sports2i*, a Korean sports statistics firm (Park, 2008). The outcome of 249 games(the total games with a definitive outcome played by the team during the period 2007–2008) are listed in Table 2.1. The Lotte Giants won 124 out of the total 249 games which means they won 49.8% of their games. The team's winning percentage was the highest against the KIA Tigers and the lowest against the SK Wyverns. To reduce the level of variables related to the seven teams against which the Lotte Giants played, we explored variables for grouping their data. The resulting groups were the KIA Tigers, the Seoul Heroes, the group LG Twins and SS Lions, finally the SK Wyverns, reducing the total number of variables to five.

The data set was divided into a training and validation sets, each containing 70% and 30%, respectively. The training subset was used build the model, whereas the validation set was used check the performance of the model. Among the variables related to the game which were used in

Table 2.1. Results of 249 games played by the team from 2007–2008

Team	Lose	Win	Total
HANHWA Eagles	19	17	36
KIA Tigers	12	24	36
LG Twins	16	17	33
DOOSAN Bears	19	17	36
SK Wyvens	27	9	36
SAMSUNG Lions	17	19	36
WOORI Heroes	15	21	36
Total	125	124	249

this study, these pertained to the batter as follows: Singles(H1), Doubles(H2), Triples(H3), Home Runs(HR), Stolen Bases(SB), Caught Stealing(CS), Sacrifice Bunts(SB), Sacrifice Flies(SF), Bases on Balls(BB), Intentional Walks(IW), Hit by Pitch(HP), Ground Ball Double Plays(GDP), Errors(ERR), Strikeouts(KK), Home/Road game(TB), TEAM, Wild Pitches(WP), Balks(BK). The data on offense events allowed by the pitcher are distinguished by adding the letter 'P' in front of each variable(P_name of the variable). For example, P_H1 means that pitcher allowed a single.

The analysis to determine which of these game variables influenced the outcome of games played by the Lotte Giants, and the relative importance of each of the variables was based on the overall data. The data were also analyzed at two levels; by separating the batter-related data from the pitcher data. The statistical analysis was performed using the SAS Enterprise Miner, and consisted of logistic regression analysis and decision tree analysis. The estimations obtained from the two analytic methods were compared with each other for predictive accuracy.

Logistic regression analysis is quite similar to standard regression analysis, except for the difference in the type of dependent variables. Due to this difference, binary variables are used in logistic regression analysis, instead of interval or ratio scale variables used in standard regression analysis. Assuming that the number of independent variables is p, the logistic regression model can be defined as follows:

$$\log\left[\frac{p(y=1)}{1-p(y=1)}\right] = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_p.$$

The extent of influence of an independent variable on the outcome of a classification can be measured through the odds ratio. All other independent variables being constant, the odds ratio can be calculated using the following formula:

$$\text{Odds ratio} = \frac{\exp[\beta_0 + \beta_1 x_1 + \cdots + \beta_i(x_i+1) + \cdots + \beta_p x_p]}{\exp[\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \cdots + \beta_p x_p]} = \exp(\beta_i).$$

A logistic regression model was selected in the regression node, and in the selection method dialog box, stepwise variable selection was chosen in the SAS Enterprise Miner.

The results of the logistic regression model are given in Table 2.2. As for coefficients in the estimated regression equation, the coefficient was the highest for H2 [Exp(1.3741) = 3.951]. All other variables being constant, the team's odds ratio increased the most, by 3.951 times, at each additional doubles. The next highest increase in the winning probability was observed with the variable HR. Among the independent variables having a negative coefficient, the winning probability decreased the most with P_HR(the home runs allowed by the pitcher). Its coefficient being Exp(−1.2637) = .283, when all other variables are constant, the winning probability dropped .283 times at each additional home

Table 2.2. Results of logistic regression analysis for overall data

Parameter	DF	Estimate	Standard error	P-value	Exp(Estimate)
Intercept	1	0.2698	0.9563	0.7778	1.310
ERR	1	−0.7693	0.3593	0.0323	0.463
H1	1	0.5852	0.1371	<.0001	1.795
H2	1	1.3741	0.2840	<.0001	3.951
HR	1	1.1002	0.3852	0.0043	3.005
P_BB	1	−0.6547	0.1560	<.0001	0.520
P_GD	1	0.7164	0.3117	0.0215	2.047
P_H1	1	−0.4922	0.1159	<.0001	0.611
P_HR	1	−1.2637	0.3662	0.0006	0.283
P_SF	1	−1.1918	0.5567	0.0323	0.304
P_WP	1	−1.0298	0.5244	0.0495	0.357

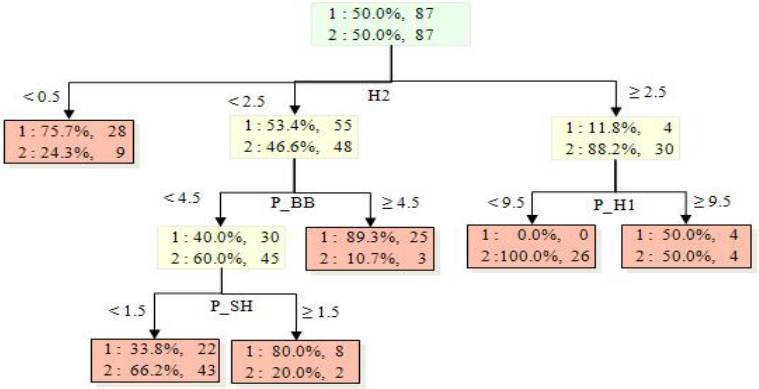

Figure 2.1. Decision tree for overall data

Table 2.3. Classification table from logistic regression model and decision tree for overall data

		Logistic regression				Decision tree			
		Lose	Win	Total	Rate	Lose	Win	Total	Rate
Result	Lose	73	14	87	0.840	84	3	87	0.966
	Win	12	75	87	0.862	29	58	87	0.667
Total		85	89	174	0.851	113	61	174	0.816

run allowed. The next highest decline in winning probability at each unit increase was observed with P_SF.

The decision tree model is a classification and prediction technique in which decision rules are placed and arranged in a diagram having a tree-like structure. The steps in a process of classification or prediction are expressed in a decision tree model(according to inference rules) that makes it easier for a researcher to understand and explain the process. The decision tree analysis was performed in this study through the following steps. Of the three candidate classification criteria in the tree node, namely CHAID(Chi-squared automatic interaction detection), CART, and C4.5, we selected the CHAID algorithm. The minimum number of observations in a leaf was set to 5; the number of observations required for a split search, to 10; and the maximum number of branches from a node, to 3. The resulting decision tree having the maximum depth of a tree of 6 is shown in Figure 2.1.

Table 2.3 compares the classifications accuracy of logistic regression and decision tree and Table

Table 2.4. Results of classification from logistic regression model and decision tree for overall data

	Training set			Validation set		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
LR	85.05%	86.20%	83.90%	84.00%	86.48%	81.58%
DT	81.61%	66.67%	96.55%	72.00%	72.97%	71.05%

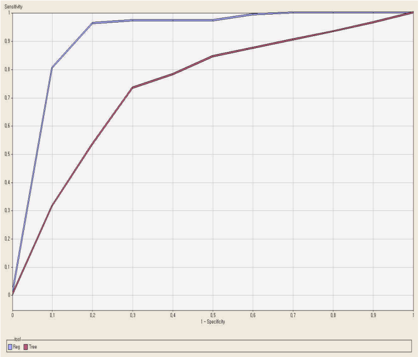

Figure 2.2. The receiving operating characteristic(ROC) chart for two different methods for overall data

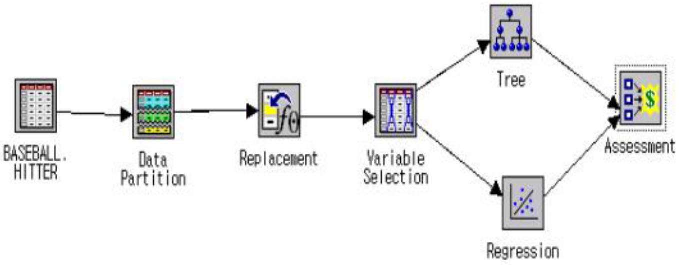

Figure 3.1. Diagram of Batter Data

2.4 shows the measurements of the accuracy, sensitivity, and specificity of two different methods. Based on the two tables, the logistic regression model outperformed the decision tree model.

Because Table 2.3 do not show the number and percentage of false positive and false negative errors, we presented a receiving operating characteristic(ROC) chart that is a technique for visualizing, organizing, and selecting classifiers based on their performance. Figure 2.2 shows the ROC chart for the two methods evaluated on data.

3. Analysis of Batter Data

To determine which of the batter-related factors influence the outcome of a game, we performed a logistic regression analysis and decision tree on the batter data using the SAS Enterprise Miner.

The analysis of the batter data using the logistic regression model took place in the following the steps illustrated in Figure 3.1. After obtaining the input variables through the above-described step, we selected the logistic regression model in the regression node. Next, stepwise variable selection was chosen in the selection method dialog box.

Table 3.1. Results of logistic regression analysis for batter data

Parameter	DF	Estimate	Standard error	P-value	Exp(estimate)
Intercept	1	−1.9526	0.4052	<.0001	0.142
H1_JEONG SU KEUN	1	0.9841	0.3084	0.0014	2.675
H2_JEONG BO MYUNG	1	2.0198	0.6478	0.0018	7.537
H2_JEONG SU KEUN	1	2.6279	0.6190	<.0001	13.844
SB_PARK KI HYUK	1	3.3685	1.1215	0.0027	29.036
SH_JEONG SU KEUN	1	6.2559	2.3338	0.0073	521.092
ERR_JEONG BO MYUNG	1	−2.6080	1.0462	0.0127	0.074

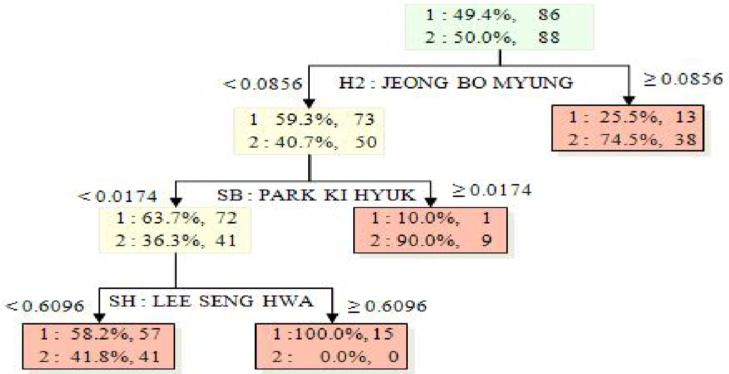

Figure 3.2. Decision tree for batter data

Table 3.2. Classification table from logistic regression model and decision tree for batter data

		Logistic regression				Decision tree			
		Lose	Win	Total	Rate	Lose	Win	Total	Rate
Result	Lose	71	15	86	0.826	47	41	88	0.534
	Win	29	59	88	0.659	14	72	86	0.837
Total		100	74	174	0.747	61	113	174	0.684

As for the coefficients of the regression equation shown in Table 3.1, the coefficient of SH_JEONG SU KEUN came to Exp(6.2559) = 521.092 and became the highest increase in the winning probability. All other variables being constant, the team's winning odds increased 521.092 times at each additional SH by the player JEONG SU KEUN. On the other hand, when concerning the independent variables having a negative coefficient, the highest decrease in winning probability was seen with ERR_JEONG BO MYUNG whose coefficient stood at Exp(−2.608) = 0.074. What this means is that all other variables being constant, the team's winning probability decreased 0.074 times at each ERR by the player JEONG BO MYUNG.

In the tree node, the Chi-square test was selected from the three candidate classification criteria, CHAID, CART, and C4.5 as in the diagram in Figure 3.2.

Table 3.2 denoted the classification table from logistic regression model and decision tree for batter data. This results showed that the logistic regression model could perform better than the decision tree with an accuracy rate of 74.7% versus 68.4%.

Table 3.3 shows the measures the accuracy, sensitivity, and specificity of two different data mining models for batter data and Figure 3.3 shows the ROC chart for two data mining models evaluated on batter data.

Table 3.3. Results of classification from the logistic regression model and the decision tree for batter data

	Training set			Validation set		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
LR	74.71%	67.05%	82.36%	54.67%	50.00%	58.97%
DT	68.39%	53.41%	83.72%	56.00%	41.67%	69.23%

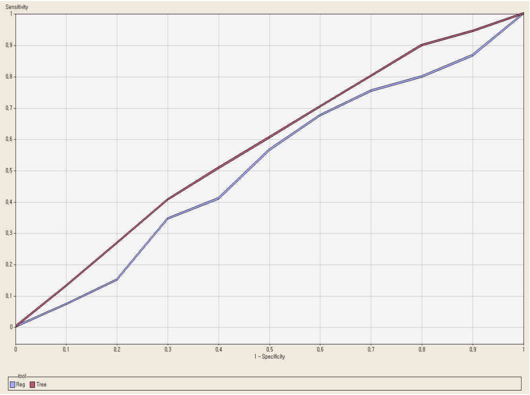

Figure 3.3. The receiving operating characteristic(ROC) chart for two different methods for batter data

Table 3.4. Results of logistic regression analysis for pitcher data

Parameter	DF	Estimate	Standard error	*P*-value	Exp(Estimate)
Intercept	1	1.148	0.4172	0.0059	3.152
SH_SONG SEUNG JUN	1	−2.4692	0.8279	0.0029	0.085
G_KIA	1	0.8198	0.3981	0.0395	2.27
G_WOORI	1	0.5357	0.3675	0.1449	1.709
G_LG	1	0.0259	0.2538	0.9188	1.026

4. Analysis of Pitcher Data

To determine which of the pitcher-related factors influenced the outcome of a game, we performed a logistic regression and a decision tree analysis on the pitcher data. The classification table providing the estimated probabilities of wins and losses(resulting from logistic regression analysis) is given in Table 3.4.

The coefficients estimated through this round of logistic regression analysis are given in Table 3.4. The coefficient of G_KIA was the highest at Exp(0.8198) = 2.27; with all other variables being constant, the team's winning probability increases 2.28 times by KIA. The next highest increase was seen with G_WOORI. With regard to the independent variables having a negative coefficient, the value of coefficient was for SH_SONG SEUNG JUN.

As for the decision tree of pitcher data, the resulting decision tree is shown in Figure 4.1. The classification results, providing the results from the logistic regression and decision tree analysis, are given in Table 4.1. When the two methods are compared in terms of classification accuracy, decision tree model outperformed the logistic regression model for pitcher data. Table 4.2 shows the measures the accuracy, sensitivity, and specificity of two different methods for pitcher data and Figure 4.2 shows the ROC chart for two data mining models evaluated on pitcher data.

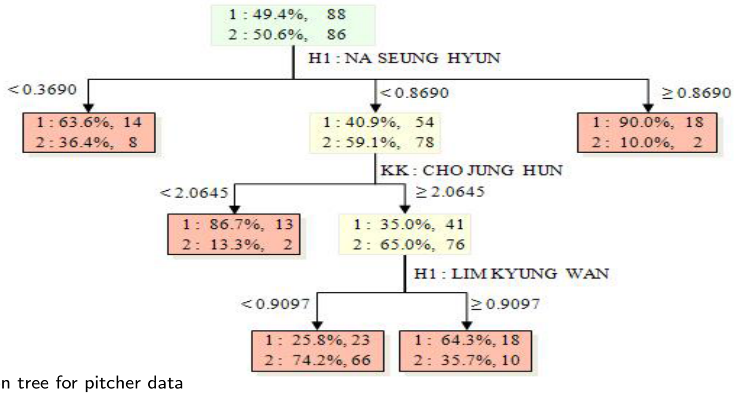

Figure 4.1. Decision tree for pitcher data

Table 4.1. Classification table from logistic regression model and decision tree for pitcher data

		Logistic regression				Decision tree			
		Lose	Win	Total	Rate	Lose	Win	Total	Rate
Result	Lose	29	57	86	0.337	63	23	86	0.733
	Win	8	80	88	0.910	22	66	88	0.75
Total		37	137	174	0.626	85	89	174	0.741

Table 4.2. Results of classification from logistic regression model and decision tree for pitcher data

	Training set			Validation set		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
LR	62.64%	90.91%	33.72%	57.33%	88.89%	28.21%
DT	74.14%	75.00%	73.26%	62.33%	69.44%	53.85%

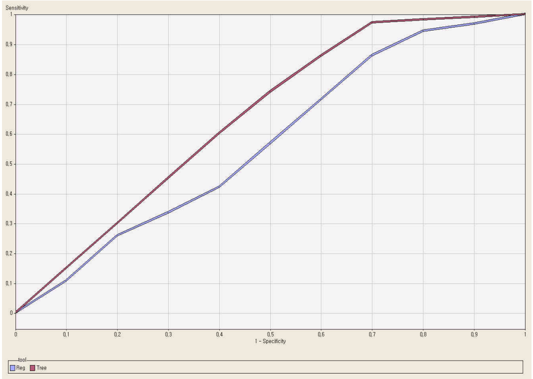

Figure 4.2. The receiving operating characteristic(ROC) chart for two different methods

5. Conclusion

In this paper, we have evaluated and contrasted decision tree classifiers with logistic regression classifiers for a real baseball dataset. This study identified general variables and variables related to individual players affecting the outcome of games played by the Lotte Giants. To achieve this goal, we separated data sets(related to batters and pitchers) and selected a series of variables to determine which of them influence the outcome of a game and to what extent they influence the outcome. Therefore, we were able to establish how much the records of individual players affect the outcome of games. When the two prediction methods are compared with regard to the overall data, logistic regression analysis was slightly superior to decision tree analysis in terms of accuracy, by 85.1% to 81.6%. With regard to the batter data, the results of logistic regression model were more accurate than those of the decision tree; however, the prediction accuracy of decision tree was outperformed the logistic regression model with regard to the pitcher data.

References

Albert, J. (1994). Exploring baseball hitting data: What about those breakdown statistics?, *Journal of the American Statistical Association*, **89**, 1066–1074.

Albright, S. C. (1993). A statistical analysis of hitting streaks in baseball, *Journal of the American Statistical Association*, **88**, 1175–1183.

Barry, D. and Hartigan, J. A. (1994). Change points in 0–1 sequences, with an application to predicting divisional winners in major league baseball, *Journal of Applied Statistical Science*, **1**, 323–336.

Bennett, J.M. (1998). *"Baseball" in Statistics in Sport,* Arnold Applications of Statistics Series, Arnold Publishing, 25–64.

Cho, Y. S. and Cho, Y. J. (2003). The research regarding a Beane Count application from Korean baseball league, *Journal of the Korean Data Analysis Society*, **5**, 649–658.

Cho, Y. S., Cho, Y. J. and Shin, S. K. (2007). A Study on winning and losing in Korean Professional Baseball League, *Journal of the Korean Data Analysis Society*, **9**, 501–510.

Choi, Y. S. and Shim, H. J. (1995). Applications of the supplementary principal component analysis for the 1982–1992 Korean pro baseball data, *The Korean Journal of Applied Statistics*, **8**, 1051–1060.

Han, J. and Kamber, M. (2001). *Data Mining: Concepts and Techniques,* Morgan Kaufmann Publishers, San Francisco.

Hong, C. S. and Choi, J. M. (2008). Steal success model for 2007 Korean professional baseball games, *The Korean Journal of Applied Statistics*, **21**, 455–468.

James, B., Albert, J. and Stern, H. S. (1993). Answering questions about baseball using statistics, *Chance*, **6**, 17–22.

Kim, H. J. (2004). Are the criteria for determining team ranking in Korean professional baseball and football reasonable from a statistical viewpoint?, *Journal of the Korean Data Analysis Society*, **6**, 1767–1775.

Lee, J. T. and Cho, H. S. (2009). An analysis on the home-field advantage in Korea professional baseball with logistic regression model, *Journal of the Korean Data Analysis Society*, **11**, 533–543.

Lee, J. T. and Kim Y. T. (2006a). A study on the estimation of winning percentage in Korean pro-baseball, *Journal of the Korean Data Analysis Society*, **8**, 857–869.

Lee, J. T. and Kim Y. T. (2006b). Estimation of winning percentage in Korean pro-sports, *Journal of the Korean Data Analysis Society*, **8**, 2105–2116.

Lee, J. T. and Kim Y. T. (2007). An effective statistical model that predicted winning percentage in Korean pro-baseball, *Journal of the Korean Data Analysis Society*, **9**, 931–942.

Lindsey, G. (1963). An investigation of strategies in baseball, *Operations Research*, **11**, 447–501.

Park, K. C. (2008). *http://www.sports2i.com.*

Shin, S. K., Park, K. C., Cho, Y. S. and Choi, S. H. (2007). A study on analyzing factors affecting the outcome of Korean professional baseball games: A case of Samsung Lions, *Journal of the Korean Data Analysis Society*, **9**, 2071–2083.