OPTIMAL REGRESSION MODEL FOR PREDICTING THE WINNING GAME AND CONTRIBUTING FACTORS IN ICE HOCKEY WORLD CHAMPIONSHIP

Eun-Jeong Lee[1] and Hye-Young Kim[1]

Korea National Sport University (KNSU), Seoul, Republic of Korea[1]

The purpose of this study was to present the statistical model to predict the winning of ice hockey game and determine the contributing factors for win in the world ice hockey championship. In order to find the optimal regression model for ice hockey, we compared two regression models (logistic and linear model) with the database of all games and the seperate databases of top/bottom teams. The logistic regression model using the seperate database was most accuratly predicted the actual outcome of games. This model and database further revealed that goalkeeping and scoring efficiencies and the number of shots on goal were significantly contributing factors to win. In addition, the results for prediction analysis of winning rate for each team indicated that offensive skills were more inportant factors than defense power to increase winning rate for teams.

KEYWORDS: ice hockey, outcome prediction, regression model

INTRODUCTION: Ice hockey is one of the most highlighted sports in the winter Olympic game. Also, the National Hockey League (NHL), the largest professional league for ice hockey, is a three billion dollar a year industry throughout United States and Canada (Roith and Magel, 2014). Each teams and countries put an effort for research to understand which factors contribute to the successful game in ice hockey as studied in other sports, such as soccer, baseball, and so on. Statistics is the most frequently used tool for sport data analysis since the large amount of data for sports game is getting easier to be accessed in these days. The logistic regression model where a dependent variable is categorical has been used to predict the outcomes of games (win or loss) in sports (Willoughby, 2002; Crowe and Middledorp, 1996) because the outcomes of each games are a binary system (win or loss) with exception of tie-game. Beside the prediction of outcome, the results of regression analysis suggest which factors significantly important for winning the game (Willoughby, 2002). The purpose of this study was to identify factors contributing to wins by finding the optimal regression model to predict the outcome of games in the ice hockey world championship tournament.

METHODS: Data were collected from every official game of the ice hockey world championship tournament from 2014 to 2017 seasons organized by the international ice hockey federation (IIHF). The official summaries for each game were provided by IIHF on the tournament website. Variables for analysis were obtained from the game summary and were listed in Table1 with definition. Shorthanded goals were excluded from variable since it is rarely occurred during the game. Tie-games are extremely rare, so only either wins or losses are included in this analysis. The JMP Pro statistical software package was employed to perform the statistical analysis.

Table 1: Definition of variables

Variables	Definition
Score difference	Difference in score (pos. value for win, neg. value for loss)
SG%	Percentage of goals from total shots; indicating a scoring efficiency
SOG	Total shots on goal; indicating an offence power
SVS%	Save as percentage of total SOG; indicating a goalkeeping ability
PIM	Penalty times in minutes
TPP	Total time of power play (PP) in minutes
PP%	Power play efficiency (PP goals/# of PP) as percentage
FOW%	Face off win (FOW) as percentage (# of FOW/total # of FO)

1. Databases

In order to find the optimal prediction model for ice hockey games, three different databases were set up to compare. First, we have the 'all team database' which includes every game in 2014~2017 seasons (N=456). The second database is the 'top team database' which includes all game from top eight teams (final rank 1~8th) of each season (N=258). The third database is the 'bottom team database' which includes all game from bottom eight teams (final rank 9~16th) of each season (N=198). The official rankings of each season were announced by IIHF at the end of the season.

2. Regression model

Two types of regression models were compared using the same database: linear regression and logistic regression model. Dependent variables (Y) were the score difference for linear model (see Table1) and "1" or "0" (win or loss) for logistic model, while both model had the same independent variables listed in Table1. The linear regression model is expressed as an equation linear in X, such as $Y(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$ (where X is independent variable). When the predicted Y through the model is positive value, the outcome prediction is determined as a win. On the other hand, if the predicted Y is negative value, the prediction is determined as a loss. For logistic regression model, dependent variable is Y€ {0,1}, with 0 indicating a loss and 1 indicating a win for the team. The model is expressed as

$$Y(X) = \frac{e^{(Z)}}{1 + e^{(Z)}} \quad (\text{where } Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$$

When the predicted probability (P) is greater than 0.5, it indicated the outcome prediction is a win, while P<0.5 indicate a loss. In both models, the predicted outcome was compared with the actual result of each game, then the prediction accuracy was calculated as a percentage of number of accurate prediction to the total number of games.

3. Team analysis

We have also performed analysis to identify important factors to increase the winning rate of the team (N=14). For each variable, the average of every games for each team were calculated and used for independent factors for the linear regression model. The total winning rates for each team during the entire seasons were calculated by (# of winning game / # of total game) × 100. The calculated winning rate of each team was used as a dependent variable. The linear regression analysis was employed in order to identify the important factors for increasing the winning rate. A simulation method was applied to estimate the correlation between specific independent variables and the winning rate by increasing the sample size. A total of 19 teams participated in the world championships during 2014~7 seasons. However, five teams who had less than 10% of the winning rate were excluded in the team analysis.

RESULTS and DISCUSSION: By using the regression model described, winning probabilities (P) for each game were calculated to predict a win (P>0.5) or loss (P<0.5). Regardless of regression model and database, the prediction accuracy was greater than 95%. As shown in table2-4, logistic regression model with the separate databases of top/bottom team presented the greatest accuracy for win or loss prediction.

Table 2: Predicted vs. actual results: All team database (logistic and linear model)

Predicted result	Actual result – Logistic model		Actual result – Linear model	
	Win	Loss	Win	Loss
P>.5 (win)	**222**	8	**224**	10
P<.5 (loss)	6	**219**	4	**217**
Total	228	227	228	227
Prediction accuracy	97.4% (222 out of 228)	96.5% (219 out of 227)	98.2% (224 out of 228)	95.6% (217 out of 227)
	Total 96.9% (441 out of 455)		**Total 96.9%** (441 out of 455)	

Table 3: Predicted vs. actual results: Top team database (logistic and linear model)

Predicted result	Actual result – Logistic model		Actual result – Linear model	
	Win	Loss	Win	Loss
P>.5 (win)	**175**	2	**176**	6
P<.5 (loss)	3	**78**	2	**74**
Total	178	80	178	80
Prediction accuracy	98.3% (175 out of 178)	97.5% (78 out of 80)	98.9% (176 out of 178)	92.5% (78 out of 80)
	Total 98.1% (253 out of 258)		**Total 96.9%** (250 out of 258)	

Table 4: Predicted vs. actual results: Bottom team database (logistic and linear model)

Predicted result	Actual result – Logistic model		Actual result – Linear model	
	Win	Loss	Win	Loss
P>.5 (win)	**48**	1	**48**	2
P<.5 (loss)	2	**146**	2	**145**
Total	50	147	50	147
Prediction accuracy	96.0% (48 out of 50)	99.3% (146 out of 147)	96.0% (48 out of 50)	98.6% (145 out of 147)
	Total 98.5% (194 out of 197)		**Total 98.0%** (193 out of 197)	

We speculates that the coefficients of variables in their regression models could be more accurately estimated when the separate database was used than using one model for all games. We could also learn from the results of the separate database whether differences are existed in games between top and bottom teams. Table5 shows the results of the logistic regression analysis (variable coefficient with the corresponding *p-values* in parentheses) with separate databases (top and bottom teams). Based on the results, goalkeeping (SVS%) and scoring (SG%) efficiencies, and the number of shots on goal (SOG) were highly significant factors to contribute for win in both database. Face-off won (FOW%) and power play efficiency (PP%) seems to be important factors especially among the top team games, however there are no statistical significances in 5% level. Comparing two database (top teams vs. bottom teams), scoring efficiency is more important than goalkeeping efficiency in the top team database, while those two factors are similarly important in the bottom team database (Table5). These results suggest that coaches ought to emphasize the scoring efficiency if teams are in top ranking for winning strategies.

Table 5: Results of logistic regression analysis (*p-value*s in parentheses)

	Database	
Variables	Top	Bottom
β_0	-118.883 (<0.0001)****	-104.245 (<0.0001)****
SVS%	0.976 (<0.0001)****	0.885 (<0.0001)****
SG%	1.205 (<0.0001)****	0.746 (<0.0001)****
SOG	0.435 (0.0003)***	0.488 (0.0004)***
FOW%	0.131 (0.13)	0.059 (0.36)
PP%	0.049 (0.15)	0.017 (0.44)
PIM	-0.086 (0.28)	-0.0004 (0.99)
TPP	0.109 (0.54)	0.014 (0.93)

Note: ****Significant at the 0.01% level; ***Significant at the 0.1% level

On the other hand, we performed the team analysis by using a linear regression analysis to determine the contributing factors for increasing the winning rate of teams throughout all four seasons (2014~2017). Table6 shows the results of regression analysis which indicate that scoring efficiency (SG%) and number of shots on goal (SOG) are significant factors to increase the winning rate of teams. Figure1 presents the results of simulation for relations between SG% and predicted winning rate (Fig1, blue) and SOG and predicted winning rate (Fig.1, red) when other factors are fixed in average values. The result of team analysis suggests that offense skills are more important than defence skills for increasing the overall winning rate during the world championship.

Table 6: Linear regression model results for team analysis: variable coefficient (*p-value*)

Variables	Results
β_0	-259.09 (0.134)
SG%	5.14 (0.003)**
SOG	2.62 (0.011)*
PP%	-1.14 (0.276)
SVS%	2.76 (0.284)
TPP	-7.93 (0.331)
FOW%	0.22 (0.824)
PIM	0.03 (0.986)

Note: ** Significant at the 1% level
* Significant at the 5% level

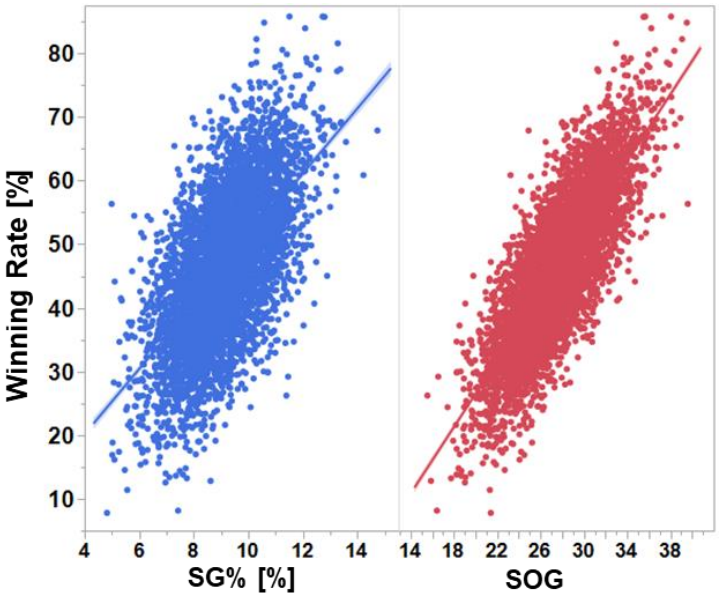

Figure1: Simulation of SG% and SOG for winning rate

CONCLUSION: These findings show that logistic regression model better predict the outcome of the games. Also, using the separate (top teams vs. bottom teams) databases increases the prediction accuracy for regression analysis. Based on regression analysis for games and for teams, the scoring efficiency (SG%) seems to be the most important factors to win a game. In sports game, regression analysis can be used as a powerful tool to identify the important factors for a winning game, and thus players and coaches are able to exploit these results for training or making a strategy for a game. Furthermore, our methodological development can be applied to the study in sports biomechanics, for example, to determine the important biomechanical factors for athletes to enhance their performance in sports.

REFERENCES

Crowe, S. & Middledorp, J. (1996). A comparison of leg before wicket rates between Australians and their visiting teams for test cricket series played in Australia. *The Statistician*, 45(2), 255-62.

Koo, D., Panday, S., Xu, D., Lee, C. & Kim, H. (2016). Logistic regression of wins and losses in Asia league ice hockey in the 2014-2015 season. *International Journal of Performance Analysis in Sports*, 16, 871-880.

Roith, J. & Magel, R. (2014). An analysis of factors contributing to wins in the national hockey league. *International Journal of Sports Science*, 4(3), 84-90.

Willoughby, K (2002). Winning games in Canadian football: a logistic regression analysis. *The College Mathematics Journal*, 33(3), 215-20.

ACKNOWLEDGEMENTS: This research was supported by Sports Science Convergence Technology Development Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2014M3C1B1034028).