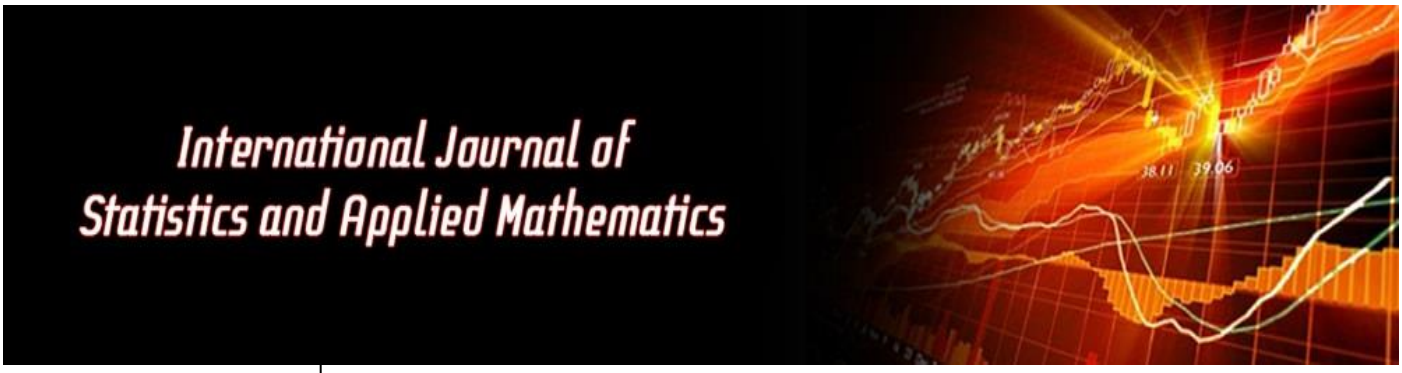

ISSN: 2456-1452
Maths 2023; 8(1): 17-21

https://www.mathsjournal.com
Received: 11-10-2022
Accepted: 15-12-2022

Prashant Kumar Choudhary
Ph.D. Scholar, Department of Sports Management, Lakshmibai National Institute of Physical Education, Gwalior, Madhya Pradesh, India

Suchishrava Dubey
Ph.D. Department of Sports Psychology, Lakshmibai National Institute of Physical Education, Gwalior, Madhya Pradesh, India

Dinesh Brijwal
Master Degree Student, Department of Sports Psychology, Lakshmibai National Institute of Physical Education, Gwalior, Madhya Pradesh, India

Rajan Paswan
Master Degree Student, Department of Physical Education Pedagogy, Lakshmibai National Institute of Physical Education, Gwalior, Madhya Pradesh, India

Corresponding Author:
Prashant Kumar Choudhary
Ph.D. Scholar, Department of Sports Management, Lakshmibai National Institute of Physical Education, Gwalior, Madhya Pradesh, India

A statistical model to predict the results of Novak Djokovic's matches in the Australian open tennis event using the binary logistic regression

Prashant Kumar Choudhary, Suchishrava Dubey, Dinesh Brijwal and Rajan Paswan

DOI: https://doi.org/10.22271/maths.2023.v8.i1a.921

Abstract
The purpose of the research was to construct a model that could forecast the probability of winning in the case of Novak Djokovic in the men's singles grand slam event of the Australian Open and to determine the relative relevance of the match data that contribute to victory. A total number of 147 matches were recorded for all nine years i.e., from 2013 to 2021, from the first round to the exit round over the years. One of the few assumptions in logistic regression is that the dependent variable must be binary in nature. Therefore, the dependent variable selected for this study was Match Outcome (Win/Loss). Ace, (DF) Double Fault, (FS) First Serve, (FSPW) first serve point win, (SSPW) second serve point win, (BPC) Breakpoint converted, and (TPW) Total point win was selected as the predictor variables. All the data were collected from ATP world tour.com. In order to accomplish the goals of the research, the only matches that Novak Djokovic competed in during the Grand Slam AO (Australian Open), were analyzed. The prediction of the likelihood of Mr Novak Djokovic winning or losing in the men's singles Australian open grand slam by fitting the logistic regression model. According to the statistical significance of the predictor variables, they were numerically weighted and can be used to predict the match outcome. Out of seven predictor variables, only the variable Breakpoint Converted was included in the prediction model with a coefficient of determination (R^2) of 0.424 (Cox & Snell) and 0.588 (Nagelkerke). The case adds seven independent variables and one dependent binary logistic variable for all the Australian Open Grand Slam matches played from 2013 to 2021. The given result of it verifies conclusive evidence that the prediction fits quite well as it classifies an 88.9% winning probability.

Keywords: Australian open, statistical model, prediction, binary logistic regression, win/loss, probability

1. Introduction
Several researchers have conducted statistical evaluations of tennis matches (Carter and Crews, 1974; Miles, 1984) [35, 36]. Assuming that player A has a constant chance of winning a point on his or her own is a typical practice. An Analysis of matches is a practice that is popular in many different sports and is seen as an essential procedure since it helps coaches to gather objective information that can be utilized to offer feedback on players' performances (Carling *et al*., 2005) [37]. As a result of the fact that coaches are prone to generating subjective judgments and may be unable to consistently remember events, a growing number of coaches are turning to match analysis as a means of enhancing the training process for their players and teams (Hughes and Franks, 2004) [38]. The primary objective of doing match analysis is to determine the areas in which one's team excels as well as those in which it could use improvement. This will, in turn, make it possible for the former to get further attention and the latter to be addressed. Similarly, a coach who is evaluating the performance of the competition will utilize the data to figure out how to exploit the vulnerabilities of the other team while minimizing the impact of the other team's strengths (Carling *et al*., 2008) [39]. In many respects, the year 2022 was an unusual one for Novak Djokovic, but 2023 is shaping up to be everything but one of those years. The twenty-first-time Grand Slam winner will definitely compete in the Australian Open at the end of the year and it is quite likely that he will have no trouble competing in the events held in the United States when the time comes. Federer has won a total of 21 Grand Slam titles.

In spite of the challenges he faced, the Serb turned in a performance that will go down in history, winning five championships during the season, including the Wimbledon Championships and the ATP Finals to round off the year. He concluded the year with the largest prize money before bonus money, having earned approximately $10 million from contests on their own. However, the 35-year-old was unable to maintain his position as the World No. 1 and finished the season in the fifth rank. When Novak Djokovic won the French Open in 2021, he established a record that has since been surpassed by Rafael Nadal when he won the Australian Open in 2022. Novak Djokovic was the first player ever to win the Double Career Grand Slam. Rafael Nadal won the Australian Open in 2022. However, the fifth-ranked player in the world has a shot at winning the Triple Career Grand Slam in the year 2023. The purpose of the research was to construct a model that could forecast the probability of winning in the case of Novak Djokovic in the men's singles grand slam event of the Australian Open and to determine the relative relevance of the match data that contribute to victory.

2. Methodology

For the purpose of the study, the last eight years of matches of the Australian Open men's tennis singles, the grand slam event was taken into consideration for analyzing probabilities of winning in the case of Novak Djokovic. A total number of 147 matches were recorded for all nine years i.e., from 2013 to 2021 from the first round to the exit round over the years. One of the few assumptions in logistic regression is that the dependent variable must be binary in nature. Therefore, the dependent variable selected for this study was Match Outcome (Win/Loss). Ace, (DF) Double Fault, (FS) First Serve, (FSPW) first serve point win, (SSPW) second serve point win, (BPC) Breakpoint converted, and (TPW) Total point win were selected as the predictor variables. All the data were collected from ATP world tour.com. In order to accomplish the goals of the research, the only matches that Novak Djokovic competed in during the Grand Slam, more specifically the Australian Open, were analyzed using statistical methodology. The prediction model was developed using a technique called binary logistic regression. In order to better understand the nature of the data, descriptive statistics were used. Before beginning the analysis, all of the presumptions were addressed and taken care of. Statistical Package for the Social Science (SPSS) version 26.0 was used in order to accomplish this goal. A value of 0.05 was chosen to represent the level of significance.

3. Results and Discussion

In contrast to linear regression and other broad linear models that are founded on ordinary least squares algorithms, logistic regression does not make nearly as many of the same fundamental assumptions. These assumptions include linearity, normalcy, homoscedasticity and measurement level. Therefore, only descriptive statistics (such as mean, standard error of the mean, standard deviation, skewness, kurtosis, etc.) were used to see the nature of the data, and the correlation matrix was used to check the assumption of high multicollinearity among the variables. This is one of the few assumptions that need to be fulfilled and therefore only descriptive statistics were used (Choudhary, 2022) [8].

Table 1: Descriptive Statistics

Descriptive Statistics							
Predictors	**N**	**Mean**	**Std. Deviation**	**Skewness**		**Kurtosis**	
	Statistic	**Statistic**	**Statistic**	**Statistic**	**Std. Error**	**Statistic**	**Std. Error**
ACE	9	8.7604	3.14010	.241	.717	1.616	1.400
Doublefault	9	2.5674	1.41871	2.072	.717	4.928	1.400
FS	9	66.67	2.236	-1.102	.717	1.300	1.400
FSPW	9	77.67	3.391	-.473	.717	-.977	1.400
SSPW	9	33.89	3.586	1.057	.717	.795	1.400
BPC	9	57.33	3.317	-.349	.717	-.525	1.400
TPW	9	46.78	8.969	-.429	.717	-1.405	1.400

Table 2: Correlations Matrix

Correlations								
		ACE	**DF**	**FS**	**FSPW**	**SSPW**	**BPC**	**TPW**
ACE	Pearson Correlation	1	-.332	.618	.199	-.787*	.108	.296
	Sig. (2-tailed)		.383	.076	.609	.012	.783	.439
	N	9	9	9	9	9	9	9
DF	Pearson Correlation	-.332	1	-.646	-.454	.562	-.732*	-.545
	Sig. (2-tailed)	.383		.060	.220	.116	.025	.129
	N	9	9	9	9	9	9	9
FS	Pearson Correlation	.618	-.646	1	.528	-.395	.405	.644
	Sig. (2-tailed)	.076	.060		.144	.293	.280	.061
	N	9	9	9	9	9	9	9
FSPW	Pearson Correlation	.199	-.454	.528	1	-.363	.522	.605
	Sig. (2-tailed)	.609	.220	.144		.337	.149	.084
	N	9	9	9	9	9	9	9
SSPW	Pearson Correlation	-.787*	.562	-.395	-.363	1	-.301	-.226
	Sig. (2-tailed)	.012	.116	.293	.337		.431	.558
	N	9	9	9	9	9	9	9
BPC	Pearson Correlation	.108	-.732*	.405	.522	-.301	1	.738*
	Sig. (2-tailed)	.783	.025	.280	.149	.431		.023
	N	9	9	9	9	9	9	9
TPW	Pearson Correlation	.296	-.545	.644	.605	-.226	.738*	1
	Sig. (2-tailed)	.439	.129	.061	.084	.558	.023	
	N	9	9	9	9	9	9	9

*.Correlation is significant at the 0.05 level (2-tailed).

In logistic regression, one of the assumptions is that there should not be a substantial degree of multicollinearity among the variables that are being analyzed independently. The multicollinearity assumption was validated with the help of the correlation matrix table that can be seen above. This table displays the correlation coefficient that was found between each set of variables. Even though there is a considerable connection between the variables, none of the variables were determined to be significantly associated. This is despite the fact that there is a correlation between the variables. This was validated by using SPSS to perform a calculation known as the Variance Inflation Factor (VIF). The VIF value was 1, which indicates that there was no multicollinearity between the independent variables. This was the case for all of the variables. As a result, we are able to proceed with the analysis of the logistic regression.

Table 3: Omnibus Tests of Model Coefficients

Omnibus Tests of Model Coefficients				
		Chi-square	**df**	**Sig.**
Step 1	Step	4.959	1	.026
	Block	4.959	1	.026
	Model	4.959	1	.026

As compared to the-2 Log Likelihood value (i.e. 124.589) of the null model, the omnibus test of model coefficients shows a significant decrease in the-2 Log Likelihood value (i.e. 6.499[a]), (Raizada, 2018) [33] which means the developed model is a significantly better fit than the null model.

Table 4: Model Summary

Model Summary			
Step	**-2 Log likelihood**	**Cox & Snell R Square**	**Nagelkerke R Square**
1	6.499[a]	.424	.588
a) Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.			

Unlike linear regression in logistic regression, there is no actual R^2 (Coefficient of Determination) value, which summarizes the proportion of variance in the dependent variable, explained by the independent variable selected by the model. The higher the proportion better will be the model. It can be seen from the above table that in the second model the value of Nagelkerke is.588 and the value of Cox & Snell R-square is found to be.424. Both Nagelkerke and Cox & Snell R-square values are the approximation of the actual value. The Nagelkerke value was considered for the developed model because in Cox & Snell R-square even for a "perfect" model with categorical outcomes, it has a theoretical maximum value of less than 1. Nagelkerke is the adjusted version of the Cox & Snell R-square that adjusts the scale of the statistic to cover the full range from 0 to 1. The value of Nagelkerke is .588 which means 58.8% of the variability in the dependent variable is explained by the selected independent variables.

Table 5: Hosmer and Lemeshow Test

Hosmer and Lemeshow Test			
Step	**Chi-square**	**df**	**Sig.**
1	3.217	6	.781

The Hosmer-Lemeshow test (HL test) is a goodness of fit test for the developed logistic regression model. It tests the null hypothesis that the fitted model is correct, which means the p-value should be insignificant to reject the null hypothesis. In the above table, the p-value is.781 which is greater than.05. Hence the model fit is good, in other words, the observed event rates match the expected event rates in population subgroups.

Table 6: Classification Table[a]

Classification Tables					
	Observed		**Predicted**		
			WIN-LOSS		**Percentage Correct**
			LOSS	**WIN**	
Step 1	Winloss	Loss	2	1	66.7
		Win	0	6	100.0
	Overall Percentage				88.9
a. The cut value is .500					

The above table shows the summary of correct and wrong classification of the subjects in match Outcome (i.e., Loss or Win) on the basis of the developed regression model. It unveils the number of wins predicted by the logistic regression model compared to the number actually observed and similarly the number of losses predicted by the logistic regression model compared to the number actually observed. Overall, 88.9% of matches were correctly classified on the basis of selected independent variables (Ma, 2013) [40].

Table 7: Variables in the Equation

Variables in the Equation							
		B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
Step 1[a]	BPC	.720	.471	2.334	1	.127	2.054
	Constant	-40.127	26.694	2.260	1	.133	.000
a. Variable(s) entered on step 1: BPC.							

The above table provides the regression coefficient (B), the Wald statistic (used to test the significance of individual coefficients in the model), and the all-important Odds Ratio (Exp (B)). "B" coefficients are also known as unstandardized coefficients and are used to develop the regression equation (Bewick, Cheek & Ball, 2005) [4]. Only the variable (BPC) Breakpoint converted is selected by the model. Tennis is a volatile sport in which the outcome cannot be decided in a short amount of time. The outcome is determined by a number of different elements, all of which interact with one another to produce an unexpected outcome. These factors are falls in crucial situations/times, default from the tennis player, unable to break the server in different point situations and also unable to convert the breakpoint, and not being able to hold the service at the crucial games, One can find many instances addressed in the game of tennis where first serve percentage and the return of the services play the paramount role in increasing the winning percentage of a player (Elliott, 2001) [15]. The stature of a player in terms of court coverage and recovery back to the center, the 25-sec short clock rule, and many more. Most of the paramount variables selected by the model are which contribute to a significant difference in the game of Novak Djokovic during the Australian Open rounds. Furthermore, the Australian Open Grand Slam stage where is where elite athletes with an almost similar level of playing ability take part on the basis of ranking provided by an association of tennis professionals. What matters is the subtle difference incurred due to the experience of handling a critical situation that too in front of a heavily packed stadium and also defending the title as well.

4. Regression equation

Regression Equation Using regression coefficients (B) of the model shown in Table 7, the regression equation was developed which is as follows: Logit = -40.127 +.720 (Team Score)

Odds = $_e$logit = e -40.127 +.720 (BPC) Breakpoint converted

$$P(Y) = \frac{\text{odds}}{1+\text{odds}}$$

The above regression equation can be used to predict the match outcome (i.e., Win/Loss) for Novak Djokovic on the basis of one predictor/independent variable (i.e. Break Point Converted) of the matches played in the 7 rounds of the Australian Open men's singles Grand slam. It will only explain 58.8% of the variability in the dependent variable, the remaining percentage of the variability (41.2%) may explain by some other variables.

5. Conclusion

Novak Djokovic is recognized for his fantastic on-court skill and the extraordinary mental toughness he demonstrates when playing tennis. He is generally considered one of the best tennis players of all time and is known for both aspects of his game. Djokovic is the epitome of the proverb "when the going gets difficult, the tough get going", which is a well-known phrase that means "when the going gets tough, the tough get going." In the 2010s decade, Djokovic won a total of 16 grand slams, which is more than Roger Federer and Rafael Nadal combined, who won a total of 15 grand slams in the decade, further cementing his dominance. In explaining his dedication to/and understanding of mindfulness in his book, 'Serve to Win', he said, "I do it every day for about 15 minutes and it is as important to me as my physical training… Instead of trying to silence your mind or find 'inner peace', you allow and accept your thoughts as they come…they do bounce around like crazy, but they're supposed to, your job is to let them come and go". One of the main reasons for this dominance is Djokovic's mental toughness, which enabled him to win more grand slams than either Federer or Nadal. Hence, it is also recommended to other players regardless of playing at the grand slam stage that Practicing yoga creates a calm state of mind, greater selective concentration, focusing and willingness to cope with frustration (Dubey, 2021) [13]. Notably, this study has its limitations, one of the major limitations to this study was the absence of Novak Djokovic's participation in the Australian Open for the year 2022 due to his holding on to the decision to stay unvaccinated. He was also barred from playing the same year's U.S. Open because of his vaccination status. Owing to the results of the study one can also assume a bit less chance of Novak winning the Australian Open title. But then it cannot be overshadowed that despite Novak's absence in the AO Australian Open and U.S Open grand slam 2022 Novak won the year-ending Nitto ATP finals which are played every year and it only constitutes the world's top 8 players. This also states in a way that Novak Djokovic's performance is impeccable and unparallel as always.

The present study was focused on the prediction of the likelihood of Mr. Novak Djokovic winning or losing in the men's singles Australian open grand slam by fitting the logistic regression model. The case adds seven independent variables and one dependent binary logistic variable for all the Australian Open Grand Slam matches played from 2013 to 2022. The given result of it verifies conclusive evidence that the prediction fits quite well as it classifies an 88.9% winning probability.

6. References

1. Assumptions of Logistic Regression-Statistics Solutions; c2018. http://www.statisticssolutions.com/assumptions-oflogistic-regression/.
2. Ballesteros C, Alexendre D. Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league Journal of Sports Science and Medicine. 2010;9(2):288-293.
3. Barnett T, Clarke SR. Combining player statistics to predict outcomes of tennis. IMA Journal of Management Mathematics. 2005;16(2):113-120.
4. Bewick V, Cheek L, Ball J. Statistics review 14: logistic regression. Critical Care. 2005;9(1):1-7.
5. Christie CJ, King GA. Heart rate and perceived strain during batting in a warm and cool environment. International Journal of Fitness. 2008;4(1):33-38.
6. Boulier B, Stekler H. Are sports seeding good predictors? An evaluation. International Journal of Forecasting. 1999;15(1):83-91.
7. Chiu FC. Comparative analysis on the skill characteristics of men's singles matches in 2008 Grand Slam tournaments. Sports and Exercise Research. 2010;12:83-95.
8. Choudhary PK, Dubey S, Singh D. A statistical model to forecast the outcome of the golden state warriors through NBA season matches. International Journal of Physical Education, Sports, and Health. 2022;9(6):134-137. https://doi.org/10.22271/KHELJOURNAL.2022.V9.I6B.2685
9. Collinson L, Hughes M. Surface effect on the strategy of elite female tennis players. Journal of Sports Sciences. 2003;21(4):266-267.
10. Corral J, Prieto-Rorriguez J. Are differences in ranks good predictors for Grand Slam tennis matches? International Journal of Forecasting. 2010;26(3):551-563.
11. Cox DR, Snell EJ. Analysis of Binary data. Second Edition. Chapman & Hall; c1989.
12. Cross R, Pollard G. Grand Slam men's singles tennis 1991-2009: Serve speeds and other related data. ITF Coaching & Sport Science Review. 2009;16(49):8-10.
13. Dubey S, Choudhary PK. A Review Paper: Effect of Stress and Anxiety in Sports Performance and Inventive Approaches to Overcome. Journal of Emerging Technologies and Innovative Research. 2021;8(12):7-14. https://doi.org/10.1123/tsp.14.4.360
14. Schultz RW. Sports and mathematics: A definition and delineation. Research Quarterly. 1980;51(1):37-49.
15. Elliott B, Saviano N. Serves and returns. In Robert P & Groppel J. (Eds.), World-class tennis technique. Champaign, IL: Human Kinetic; c2001. p. 207-222. https://www.researchgate.net/publication/235785628_Winning_matches_in_Grand_Slam_men's_singles_An_analysis_of_player_performance-related_variables_from_1991_to_2008 [accessed Dec 18 2022].
16. Erricsson KA, Krampe RT, Tesch-Romer C. The role of deliberate practice in the acquisition of expert performance. Psychological Review. 1993;100(3):363-406.
17. Field A. Discovering statistics using SPSS (3rd ed.). London: Sage; c2009.

18. Gilsdorf K, Sukhatme V. Testing Rosen's sequential elimination tournament model: Incentives and player performance in professional tennis. Journal of Sports Economics. 2007;9(3):287-303.
19. Gould D, Petlinchkoff L, Simons J, Vevera M. Relationship between competitive state anxiety inventory-2 subscale scores and pistol shooting performance. Journal of Sport Psychology. 1987;9(1):33-42.
20. Harris N. Exclusive: Djokovic, Nadal, Federer-as close to perfection as tennis has ever been; c2012.
21. Helsen WF, Starkes JL, Hodges NJ. Team sports and the theory of deliberate practice. Journal of Sport and Exercise Psychology. 1998;20(1):12-34.
22. Hornery DJ, Farrow D, Mujika I, Young W. An integrated physiological and performance profile of professional tennis. British Journal of Sports Medicine. 2007;41(8):531-536.
23. Jones G. More than just a game: Research, developments and issues in competitive anxiety in sport. British Journal of Psychology. 1995;86(4):449-479.
24. Kirkendall DT, Garrett WE. The effects of aging and training on skeletal muscle. The American Journal of Sports Medicine. 1998;26(4):598-602.
25. Klaassen F, Magnus J. Forecasting the winner of a tennis match. European Journal of Operational Research. 2003;148(2):257-267.
26. Locke EA, Latham GP. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. American Psychologist. 2002;57(9):705-717.
27. Loffing F, Hagemann N, Strauss B. The serve in professional men's tennis: Effects of players' handedness. International Journal of Performance Analysis in Sport. 2009;9(2):255-274.
28. Shang-Min Ma, Chao-Chin Liu, Tan Yue, Shang Ma. Winning matches in Grand Slam men's singles: An analysis of player performance-related variables from 1991 to 2008. Journal of Sports Sciences. 2013;31(11):1147-1155. 10.1080/02640414.2013.775472.
29. MacCurdy D. Talent identification around the world and recommendations for the Chinese Tennis Association; c2006. Retrieved from http://www.itftennis.com/shared/medialibrary/pdf/original/IO_18455_original.pdf.
30. O'Donoghue PG, Brown E. The importance of service in Grand Slam singles tennis. International Journal of Performance Analysis in Sport. 2008;8(3):70-78.
31. O'Donoghue PG, Ingram BA. Notational analysis of elite tennis strategy. Journal of Sports Sciences. 2001;19(2):107-115.
32. Pluim B. Medical considerations when identifying talent. ITF Coaching & Sport Science Review. 2006;39:6-7.
33. Raizada Shiny, Bagchi Amritashish, Menon Harishankar, Nimkar Nayana. Predicting the outcome of ICC cricket world cup matches; c2018.
34. Sporting Intelligence. Retrieved from http://www.sportingintelligence.com/2012/02/13/exclusive-djokovic-nadal-federer-asclose-to-perfection-as-tennis-has-ever-been-130201/
35. Carter Jr WH, Crews SL. An analysis of the game of tennis. The American Statistician. 1974;28(4):130-134.
36. Miles RE, Snow CC. Designing strategic human resources systems. Organizational Dynamics. 1984;13(1):36-52.
37. Kahn BB, Alquier T, Carling D, Hardie DG. AMP-activated protein kinase: ancient energy gauge provides clues to modern understanding of metabolism. Cell Metabolism. 2005;1(1):15-25.
38. Hughes M, Franks IM, editors. Notational analysis of sport: Systems for better coaching and performance in sport. Psychology Press; c2004.
39. Carling C, Bloomfield J, Nelsen L, Reilly T. The role of motion analysis in elite soccer. Sports Medicine. 2008;38(10):839-62.
40. Ma WX. Bilinear equations, Bell polynomials and linear superposition principle. In Journal of Physics: Conference Series IOP Publishing. 2013;411(1):1-11.