ISSN 2618-1576 Vol. 6, Issue 4: 107-113, DOI: 10.25236/FSR.2024.060416

Research and analysis of winning factors in men's singles tennis based on logistic regression modelling

Fuqian Wang[1,a], Yirong Wang[2,b], Li Liao[1,*]

[1]Institute of Physical Education, Xiangnan University, Chenzhou City, Hunan Province, China
[2]School of Computer and Artificial Intelligence, Xiangnan University, Chenzhou City, Hunan Province, China
[a]2820419479@qq.com, [b]2832454707@qq.com
[*]Corresponding author: liaoligoogl@gmail.com

Abstract: *The Wimbledon Tennis Championships, the historic Grand Slam tennis tournament, is known for its grass courts and plain white dress code. Since its inaugural edition in 1877, Wimbledon has become a coveted place of honour for players and has witnessed countless legends. This study analyses the data of 128 men's singles matches published on the official website in 2023 through the literature method, and uses the P-S method for correlation analysis, and explores the factors affecting the results of the matches through the S-W and K-W tests, the independent samples t-test, the logistic regression model, and the ROC test. The results of the study showed that winning the 1st serve point had a significant effect on the match outcome, providing new training directions for players and coaches and emphasising the importance of the serve point in tennis. Based on this finding, players and coaching teams can pay more attention to the competition for serve points and improve the ability to win serve points through targeted training.*

Keywords: *winning factors, men's singles, Wimbledon Open*

1. Introduction

Tennis, as a global sport, attracts countless spectators and participants with its unique glamour and competitiveness. The Wimbledon Championships, one of the four Grand Slam tournaments in tennis, adds a unique dimension to the sport with its long history and traditional grass courts. Each year, Wimbledon brings together the world's top tennis players who compete for honours and demonstrate excellence in technique and strategy. However, through a careful review of the existing literature, we found that although research on the winning factors of tennis matches has made some progress, the singularity of the research methodology may lead to limitations of the research results, so that there are still some controversies and divergences in the research results, and the exploration of the winning factors is often confined to a match between a champion and a runner-up, which is too monotonous, and it may be difficult to accurately assess the generality of the treatment effect due to the limited number of research subjects. It may be difficult to accurately assess the prevalence and durability of treatment effects due to the limited number of subjects[1] . The aim of this study was to fill this research gap by examining the validity of data on key match-winning factors through data from 128 men's singles matches and using a variety of tests. The use of technology in tennis has focused on improving the experience and performance of players[2] , serve speed, first serve success rate, etc. To reveal the winning factors in the men's singles match at Wimbledon 2023 and to explore how these factors interact with each other to influence the final outcome of the match.

Through this study, we hope to provide tennis players and coaches with more targeted training advice, as well as contribute new perspectives and methods to the scientific study of tennis. In the next sections, we will present the details of the research methodology, the results of the data analyses, and our conclusions. We believe that this study will not only be of great significance to tennis professionals, but also provide references and insights for a wide range of sports science research.

2. Materials and methods

2.1 Research Objectives

The 128 men's singles matches and 15 indicators of their players announced for the 2023 Wimbledon Open were used for the study.

2.2 Research methodology

2.2.1 Literature method

Read the related literature and books about the influence of tennis winning factors and other relevant literature and books through the library. Collecting and analysing data through Wimbledon Championships to obtain the relevant data of the tournament and provide sufficient theoretical basis for this study.

2.2.2 Mathematical and statistical methods

ExceL was used to classify the 15 variables of ACEs, double faults, first serves, first serve percentage, second serve percentage, break success rate, unforced errors, distance covered, distance/points covered, age, height, racket handler, weight, number of tournament appearances, and time spent in the tournament as published on the official website of the 2023 Wimbledon Open. The main technical indicators affecting men's singles winning at Wimbledon 2023 were collated, analysed and statistically analysed using logistic regression model using spss2006.

2.2.3 Data sources and indicators

Through the official website of Wimbledon Open tennis tournament website to collect the required research data indicators, consult the relevant literature to select the corresponding indicators, to determine the research indicators of this research. From Table 1, the statistical indicators are as follows: ACE (X1), double faults (X2), first serve (X3), first serve scoring rate (X4), second serve scoring rate (X5), break success rate (X6), unforced errors (X7), distance covered (X8), distance/points covered (X9), age (X10), height (X11), racket handiness (X12), weight (X13), number of tournaments (X14), tournament time (X15) and so on. tournaments (X14), time spent in tournament (X15) and other 15 variables were categorised.

Table 1: 2023 Wimbledon Men's Singles Winning Factor Variable Statistics

coding	norm	Indicator unit
X1	ace	number of individuals
X2	double fault (in tennis)	number of individuals
X3	First serve.	per cent
X4	Probability of winning the 1st serve	per cent
X5	Probability of winning the 2nd serve	per cent
X6	Winning Break Points	per cent
X7	unforced error	number of times
X8	Coverage distance	surname Mi
X9	Coverage distance/point	surname Mi
X10	(a person's) age	
X11	(a person's) height	CM
X12	Licence Holding	Right 1 Left 2
X13	weight	KG
X14	Participation in tournaments	number of times
X15	competition duration	

3. Results

3.1 Correlation analysis of pre-selected variables for each indicator

Pearson's correlation coefficient measures the linear relationship between two variables by calculating the ratio of their covariance to their respective standard deviations. If the correlation coefficient is close to 1 or -1, it indicates that there is a strong linear relationship between the two variables; if it is close to 0, it indicates that there is almost no linear relationship[3] . In this study, when the absolute value is taken to be greater than 0.85, it indicates that there is a strong linear correlation between the two variables, From Table 2 , it can be seen that there is no absolute value greater than 0.85 for the 15 variables, so there is no strong correlation coefficient between the 15 variables, and all of them are entered into the regression model and included in the next step of the analysis.

Table 2: Correlation analysis

norm	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15
X1	1														
X2	.206**	1													
X3	.048	-.221**	1												
X4	.525**	.027	-.026	1											
X5	.082	-.294**	.081	.293**	1										
X6	.084	.023	.046	.147*	.089	1									
X7	.210**	.356**	-.033	-.233**	-.272**	-.034	1								
X8	.074	.093	.036	-.138*	-.041	.042	.581**	1							
X9	-.315**	-.142*	-.049	-.248**	-.122	.019	.206**	.645**	1						
X10	-.043	-.087	.032	-.029	-.049	.083	-.070	.078	.145*	1					
X11	.500**	.077	.114	.399**	.029	.068	.026	-.096	-.269**	.038	1				
X12	-.004	.019	-.009	-.059	-.017	.015	.037	-.025	-.012	.023	-.129*	1			
X13	.365**	-.027	.073	.251**	.024	.026	.024	-.094	-.268**	.117	.692**	-.061	1		
X14	-.022	-.143*	-.061	.159*	.060	.097	-.202**	.067	.211**	.744**	.047	.039	.081	1	
X15	.318**	.200**	.030	.032	.035	.029	.574**	.830**	.198**	.012	.094	-.077	.096	.006	1

3.1.1 Test for normal distribution of independent variables

In order to ensure the accuracy and reliability of the prediction of the winning factors, the research methodology of this paper emphasises the use of the final calculations of the Shapiro-Wilk Test as a measure, and the 1972 study proposed a modified version of the Shapiro-Wilk Test that could be used for large samples and used coefficients that depended on the expectation of the normal order statistic, and these coefficients are generally available, thus increasing the sensitivity of the test[4] . From Table 3, the aim of this statistical test is to ensure that the data of our selected variables conform to the properties of a normal distribution by performing a rigorous normal distribution test on 15 variables selected from the 2023 Wimbledon Open.The Shapiro-Wilk Test, which is widely used in a number of fields due to its sensitivity and accuracy, is effective in checking that a sample of data obeys a normal distribution and give the P-value as the basis for judgement. At the operational level, we will follow the criterion of P>0.05 to select variables that conform to normal distribution. This means that only if the variable corresponds to a P-value greater than 0.05 will we include it in further analyses and discussions. This criterion is based on the significance level set in statistics, which ensures that the variables we select are statistically significant, thus increasing the robustness and reliability of the findings.

Table 3: Variable tests

normality test				
norm	S-W	df	Sig.	Whether or not it conforms to a normal distribution
X1	0.928	256	0	clogged
X2	0.899	256	0	clogged
X3	0.994	256	0.481	be
X4	0.993	256	0.293	be
X5	0.985	256	0.008	clogged
X6	0.955	256	0	clogged
X7	0.964	256	0	clogged
X8	0.918	256	0	clogged
X9	0.96	256	0	clogged
X10	0.962	256	0	clogged
X11	0.965	256	0	clogged
X12	0.387	256	0	clogged
X13	0.967	256	0	clogged
X14	0.773	256	0	clogged
X15	0.963	256	0	clogged

3.1.2 Non-parametric tests of independent variables

Non-parametric tests can be used to determine whether there is a significant difference in the pre-selected variables between the two groups of winners and losers. When the data does not conform to a normal distribution, traditional parametric testing methods may not yield accurate results, whereas non-parametric tests can provide more robust inferences. In addition, the K-W test can be used to explore outliers or errors in the data, as these extreme values may affect the overall distributional characteristics of the data, and hence the results of statistical analyses based on assumptions about a particular distribution[5] . Using non-parametric tests, we can analyse the effect of each variable on the outcome separately, thus assessing more precisely the independent role of each variable in the judgement of winners and losers.

Table 4: Kruskal-Wallis Test

variant	asymptotic significance	difference
X1	0.084	insignificant
X2	0.153	insignificant
X5	0.933	insignificant
X6	0.961	insignificant
X7	0.034	statistically significant
X8	0.00	statistically significant
X9	0.00	statistically significant
X10	0.831	insignificant
X11	0.796	insignificant
X12	0.695	insignificant
X13	0.162	insignificant
X14	0.81	insignificant
X15	0.00	statistically significant

From Table 4, it can be seen that when p-value is less than 0.05,the intergroup variability of the variables is significant and when p-value is more than 0.05 the intergroup variability term of the variables is not significant. So X7 (unforced errors), X8 (distance covered), X9 (distance/points covered) and X15 (time spent in the game) are significant correlates that affect the winners and losers of that game.

3.1.3 Independent samples t-test

In the study of sports games, especially in technical statistics like serving, the researcher may be concerned about the difference between the probability of succeeding in the first serve and the probability of winning the first serve point. The independent samples t-test is a commonly used method of statistical analysis, which is mainly used to assess whether the difference between two sets of independent data is statistically significant or not, and through the independent samples t-test, we can make a comparison between these two sets of data that conform to normal distribution. Comparison, to explore whether the difference between them is significant, so as to determine whether these two variables have a significant impact on the winning factors of the game and the extent of the impact.

Table 5: Independent samples t-test

		variance equation test			t-test for the mean equation					
									95% confidence interval for the difference	
		F	Sig.	t	df	Sig. (bilateral)	mean value difference	standard error value	lower limit	limit
X3	Assuming equal variances	0.321	0.572	-1.87	254	0.063	-0.01488	0.00796	-0.03055	0.00079
	Assuming that the variances are not equal			-1.868	252.373	0.063	-0.01488	0.00796	-0.03056	0.00081
X4	Assuming equal variances	2.724	0.1	-9.523	254	0	-0.08845	0.00929	-0.10674	-0.07016
	Assuming that the variances are not equal			-9.496	246.957	0	-0.08845	0.00931	-0.10679	-0.0701

As can be seen from Table 5, the p-value of the variance chi-square test for variable X3 and variable X4 are both greater than 0.05, and the assumption of variance chi-square is valid, and checking the p-value of each variable when the variances are equal shows that the p-value of the variable X3 is greater than 0.05, and the p-value of X4 is less than 0.05, which indicates that there is a significant difference between the variables, and suggests that X4 (the probability of winning the 1st serve point) has a significant impact on the outcome of winning the game. .

3.2 Analysis of Winning Factors in Players' Competitions

As can be seen from Tables 3 and 4, the factors that produce a winning factor on the outcome of the match are X4 (probability of winning the 1st serve point), X7 (unforced errors), X8 (distance covered), X9 (distance/points covered), and X15 (time spent in the match), which are in line with the characteristics of the variability, and are therefore cited as candidate variables in the construction of the Logistic regression model.

3.2.1 Logistic regression modelling

Binary Logistic Regression is a statistical method for analysing the relationship between a binary dependent variable (with only two categories, e.g. win/lose, yes/no) and one or more independent variables (which can be quantitative or qualitative). This type of regression model is particularly useful for predicting the probability of an event occurring, such as predicting the likelihood of a player winning in a tennis match.

Table 6: Binary Logistic Regression

	B	S.E,	Wals	df	Sig.	Exp (B)
Probability of winning 1st serve point	19.257	2.679	51.663	1	0	2.31E+08
unforced error	-0.006	0.016	0.157	1	0.692	0.994
Coverage distance (metres)	0.001	0.001	2.166	1	0.141	1.001
Coverage distance points (metres)	0.032	0.115	0.077	1	0.781	1.032
competition time	-21.036	14.286	2.168	1	0.141	0
constant	-14.591	2.541	32.96	1	0	0

As can be seen in Table 6, the test value for the probability of winning the 1st serve point is 0 (omitting the number after the decimal point), which is statistically significant; the test value for unforced errors is 0.692, which is not statistically significant; the test value for the distance covered (in metres) is 2.166, which is not statistically significant; the test value for the distance point covered (in metres) is 0.0771, which is not statistically significant; the test value for the match time taken is 0.141, which is not statistically significant. test value is 0.141, which is not statistically significant.

3.2.2 ROC

The area under the ROC (Receiver Operating Characteristic) curve is an important tool for evaluating the performance of a model in a binary classification problem by plotting the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) for different thresholds to reflect the classification ability of the model. False Positive Rate (FPR) at different thresholds to reflect the classification ability of the model.The area under the ROC curve (AUC) is a commonly used metric to assess the discriminative ability of predictive models. The closer the AUC value is to 1, the stronger the discriminative ability of the model is. This quantitative approach helps to scientifically assess and compare the performance of different models[6] .

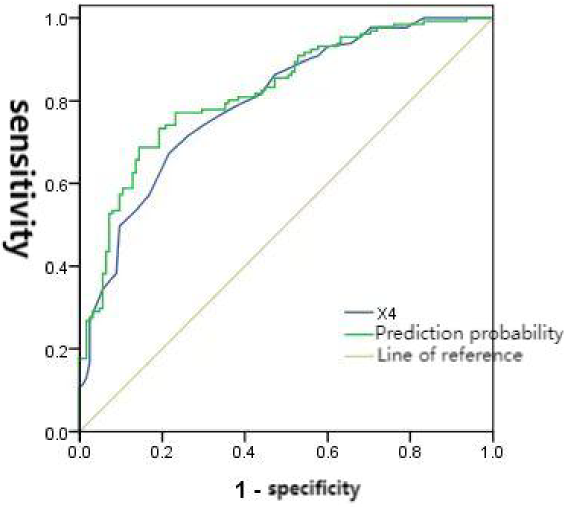

Figure 1: ROC

Table 7: ROC curve test results

Test Outcome Variables	area (of a floor, piece of land etc)	Standard errorsa	Progressive Sig.b	Approaching the 95 per cent confidence interval	
				lower limit	limit
X4	0.798	0.027	0	0.745	0.851
predictive probability	0.822	0.026	0	0.771	0.872

As can be seen in Figure 1, the Area Under the Curve (AUC) test result is 0.798, which is a very important indicator. The closer the AUC is to 1, the better the classification performance of the model.The p-value is the key statistic to measure the significance of this result, and here P=0.000 indicates that the predictive probability variable is extremely statistically significant.

Based on the test results in Table 7, we can conclude that the selected predictive probability variable, X4 (the probability of winning the 1st serve point), exhibits superior diagnostic accuracy in predicting winners and losers, and its predictive effect is significant and statistically significant, so it can be effectively applied to win/loss prediction in real-life scenarios.

4. Discussion

The probability of winning the first service point in tennis matches such as Wimbledon has a significant impact on winning or losing a match. Firstly, winning the serve is seen as an important psychological advantage in the match as it allows the player to control the pace of the match and be the first to put pressure on the opponent. In addition, analysed on a technical level, the server can take advantage of his opponent's starting position to make a tactical layout or to pose a disadvantage to him. According to On the Probability of Winning with Different Tournament Procedures (1963), the probability of each player winning a match in a competition is closely related to their performance in the early stages of the match[7] . This suggests that taking a lead at the start of a match (e.g. winning the serve) may increase the probability of winning. In addition, it is mentioned in Fluctuations of martingales and winning probabilities of game contestants (2012) that each player in a match has a certain probability of winning and this probability varies as the match progresses. Therefore, winning the first service point is not only a psychological advantage, but may also be a boost to the actual probability of winning[8] .

Based on the research that has been done, we can expect that the team that wins the first serve point will have a higher probability of winning in most cases. For example, one study found that in football, the average probability of a team scoring the first serve winning a match was 71.17 per cent. Although this is data for football, a similar trend may be observed in tennis[9] .

This study focused on the men's singles event at Wimbledon in 2023, and while the conclusions and analyses drawn from this particular tournament venue and gender range of constraints cannot be fully and directly applied to the other three Grand Slam events (e.g., the Australian, French, and U.S. Opens), it is nonetheless of specific informative value.

References

[1] D.E. Deitz, D. Cullinan et al. "Considerations for Evaluating Single-Subject Research." Remedial and Special Education (RASE) (1983).52-60.
[2] Shuaishuai Zhang, Gang Chen et al. "The Interplay Between Table Tennis Skill Development and Sports Performance: A Comprehensive Review. "Pacific International Journal(2023).
[3] Jeremy Adler and I.Parmryd. "In support of the Pearson correlation coefficient." Journal of Microscopy (2007).
[4] S.Shapiro,R.S.Francia. "An Approximate Analysis of Variance Test for Normality."Journal of the American Statistical Association (1972).
[5] S.Shapiro,M.Wilk. "An Analysis of Variance Test for Normality (Complete Samples)." Biometrika (1965).
[6] A. Janssens and FK Martens. "Rethinking modern methods: revisiting regions under the ROC curve ......" International Journal of Epidemiology (2020).
[7] D.Searls. "On the Probability of Winning with Different Tournament Procedures." Journal of the American Statistical Association(1963).1064-1081.
[8] D.Aldous and Mykhaylo Shkolnikov. "Fluctuations of martingales and winning probabilities of game contestants."arXivLabs(2012).1-17.
[9] W.Leite. "The influence of the first goal on the final result of the football match." Social Psychology(2015).29-35.