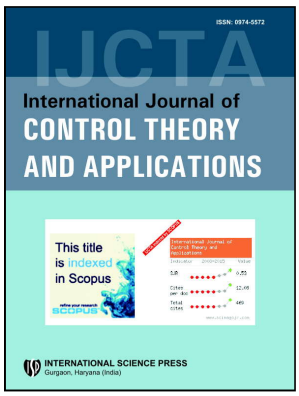

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 18 • 2017

Application of Multiple Linear Regression Models in the Identification of Factors Affecting the Results of the Chelsea Football Team

Margarita Castillo Ramirez[1], Amelec Viloria[1], Alexander Parody Muñoz[2] and Heidi Posso[3]

[1] Facultad de Ingeniería Universidad de la Costa Cl. 58 #55-66, Barranquilla - Colombia,
Emails: mcastill23@cuc.edu.co, aviloria7@cuc.edu.co
[2] Subsistema de Investigacion (SIDI) Universidad Metropolitana Calle 76 #42-78, Barranquilla - Colombia,
Email: Parody.alexander@gmail.com
[3] Programa de Bacteriologia Universidad Metropolitana Calle 76 #42-78, Barranquilla - Colombia,
Email: heidipossom@gmail.com

Abstract: The performance of the soccer teams is influenced by the technical and physical characteristics of the players, generating that the teams are measured according to these particularities. This relationship between the quality of the players and the performance of the team is known, but it would be of interest to identify the influence of a player on the result of a match, by means of the generation of a multiple linear regression model, where it is taken as Independent variables the minutes in the field of each player and as a dependent variable the points achieved by the team in the match, thus knowing that players have a significant relationship with the result, this analysis is complemented with the generation of other regression models that relate variables Associated to the behavior of the team as a collective and the result of the match, achieving a conclusion that integrates the individual and collective aspect of the players.

Keyword: Multiple regression model, analysis of performance in soccer teams, behavior patterns, competition dynamics

1. INTRODUCTION

Soccer has ceased to be for decades a game aimed solely for the recreational development of free time, has become an entire industry that moves millions of dollars at any moment, where players acquire salaries well above the average of any worker and Football clubs have been established as major assets of business consortia and it is due to this internalization of business thinking in clubs, which has brought with it that continuous improvement methodologies that previously were only thought to apply to productive processes, are now used To improve the performance of the players on the field and therefore in a better performance of the soccer team. The technology of information acquisition, processing, storage and analysis is already applied

today in the understanding of soccer as a complex system in search of a goal [1], a goal that obviously is not always achieved due to the difficulty that technicians And players to understand and assimilate those patterns of individual and collective behavior associated with good team performance, so that constant search for patterns or guides to not only understand football as a complex experimental process, but also seeks better results for The club [2].

But many of these studies analyze factors associated with individual players [3], [4] or the team as a whole [5] - [10], but so far no attempt has been made to integrate these analyzes in order to obtain A more global vision in relation to the factors that can influence the result of the football match, thus providing the main support or support for the generation of this research, since it seeks to integrate these two types of analysis into the understanding Performance of the Chelsea team of England.

2. THEORETICAL FOUNDATION

The application of statistical tools for the improvement of processes is not something new in the management of different systems, but it is the emergence of new technologies for the capture, storage, processing and analysis of information, which has allowed the boom Of data analysis on a global scale extending to unusual fields of application, through projects in Big Data or Analytics in different fields of study (Business Analytics, Sport Analytics, etc.), but it is precisely the Sport Analytics trend Which has served as a philosophical support for the research project, originated in the 1980s due to the contribution of Bill James to the statistical study of player performance in Baseball matches through his baseball writings titled Baseball Abstracts, but it is in the last decade with the advancement of technologies for information management, which has massified the use of statistical analysis to explain not only the performance of athletes, but also to predict the behavior of athletes when Generate changes in the factors that influence them, that is to say, that same approach in process optimization applied at industrial level to use it in athletes in search of the optimization of efforts in order to achieve more and better results.

The theoretical basis applied in the research focuses on the application of multiple linear regression models, which had their origin in the least squares studies devised by Legendre in 1805, which focus on explaining the behavior of one (or several) Dependent variables from a series of independent variables related through a linear equation with a series of coefficients equal to the number of independent variables statistically significant, this equation or linear model must comply with the conditions associated with the analysis of the residues or errors of the Model: the forecast error of the model must follow the behavior of a normal distribution with average equal to zero (or very close to this), the variability of forecast error must be uniform in relation to the different parameters of interest of the model, must Independence of the explanatory variables of the model.

Although the trend of Sport Analytics is new, (bringing with it the generation of new companies focused on this type of analysis and the diversification of others) the theoretical foundations associated to the research project are of extensive application and use in statistical studies of diverse nature.

3. METHODOLOGY

The study is of a correlational type, where, with the help of the ESPN sports channel ESPN (espndeportes.com), the information was taken day after day of the Chelsea football club team of the city of London in England during the 2014 season / 2015 (a total of 38 matches played) of English Premier League football, known worldwide as the Premier League, in relation to the following variables:

Minutes played by each player entered by the team

Percentage of ball possession

Shots to the arc, shots outside the arc, corners shots and close shots, of both Chelsea and its rivals.

Fouls committed, yellow cards and red cards received.

Committed offside.

Number of nationalities different from the titular players in the field.

Average age of the titular players in the field.

Team and rival training.

Condition of place or visitor.

Average position of the players by sector of the court.

Points obtained by the result of the match.

Once the information was obtained, the information from the first 28 disputed dates was used to generate the regression models that explain the behavior of the points obtained in the matches with the statistical software Statgraphics XVI, and the remaining 10 dates will be used for Test the efficiency of the model to predict the behavior of the equipment.

The multiple regression models that were generated are:

Model that relates the minutes played by each registered player and the points achieved by the team in the match.

Percentage of ball possession, bow shots, off-court shots, corner shots and close-off shots by both Chelsea and Of their rivals, fouls committed, yellow cards and red cards received, games played, number of nationalities different from the players in the field, average age of the players in the field, team and rival formation, condition Place the visitor.

Model that correlates the average positional behavior of the players per game and the points obtained in the same.

The analysis of the results of the three models of multiple linear regression that allow to have a study with an individual and collective approach of the soccer team in function of the points achieved by the disputed party [11].

4. RESULTS

The regression model associated with the number of minutes played by the players and the result of the match, presented the following series of players per position that showed correlation with the result of the match:

Table 1
ANOVA tables of the regression model associated with goalkeepers.

Source	*Sum of Squares*	*Gl*	*Squared Mean*	*Reason-F*	*P-value*
Model	149,344	2	74,672	67,75	0,0000
Residue	28,656	26	1,10215		
Total	178,0	28			
Variable		*Estimation*	*Standard Error*	*Statistical T*	*P-value*
Thibaut Courtois		0,0236978	0,00242898	9,75627	0,0000
Petr Cech		0,0337486	0,00547692	6,16197	0,0000

R-square (adjusted for g.l.) = 83.282 percent

POINTS OBTAINED = 0.0236978 * (Courtois minutes) + 0.0337486 * (Cech minutes)

Table 2
ANOVA tables of the regression model associated with the defenses

Source	*Sum of Squares*	*Gl*	*Squared Mean*	*Reason-F*	*P-value*
Model	153,408	2	76,704	77,98	0,0000
Residue	24,5919	25	0,983678		
Total	178,0	27			
Variable		*Estimation*	*Standard Error*	*Statistical T*	*P-value*
Cesar Azpilicueta		0,0270789	0,00249295	10,8622	0,0000
Filipe Luis		0,021897	0,00430342	5,08828	0,0000

R-square (adjusted for g.l.) = 85.6317 percent

POINTS OBTAINED = 0.0270789 * (minutes Azpilicueta) + 0.021897 * (minutes F. Luis)

Table 3
ANOVA tables of the regression model associated with the flyers

Source	*Sum of Squares*	*Gl*	*Squared Mean*	*Reason-F*	*P-value*
Model	152,64	2	76,3202	78,25	0,0000
Residue	25,3595	26	0,975366		
Total	178,0	28			
Variable		*Estimation*	*Standard Error*	*Statistical T*	*P-value*
Eden Hazard		0,0140432	0,0054014	2,59993	0,0152
Nemanja Matic		0,0139743	0,0054288	2,5741	0,0161

R-squared (adjusted for g.l.) = 85.2051 percent

POINTS OBTAINED = 0.0140432 * (Hazard minutes) - 0.0139743 * (Matic minutes)

Table 4
ANOVA tables of the regression model associated with the forward

Source	*Sum of Squares*	*Gl*	*Squared Mean*	*Reason-F*	*P-value*
Model	134,714	3	44,9047	25,94	0,0000
Residue	43,2858	25	1,73143		
Total	178,0	28			
Variable		*Estimation*	*Standard Error*	*Statistical T*	*P-value*
Diego Costa		0,022942	0,00352104	6,51569	0,0000
Didier Drogba		0,0270076	0,00815408	3,31216	0,0028
Loïc Rémy		0,0286303	0,00895043	3,19876	0,0037

R-square (adjusted for g.l.) = 73.7367 percent

POINTS OBTAINED = 0.022942 * 1 (Coast minutes) + 0.0270076 * (Drogba minutes) + 0.0286303 * (Rémy minutes)

Regression model that correlates the variables associated with the behavior of the team during the match and the points obtained by it:

Table 5
Regression model ANOVA tables associated with behavioral variables of the team

Source	*Sum of Squares*	*Gl*	*Squared Mean*	*Reason-F*	*P-value*
Model	166,952	5	33,3903	69,51	0,0000
Residue	11,0484	23	0,480367		
Total	178,0	28			

Variable	*Estimation*	*Standard Error*	*Statistical T*	*P-value*
% POSSESSION	0,026012	0,00659262	3,94563	0,0006
ARCHERY SHOTS	0,702837	0,110093	6,38401	0,0000
Stuck	0,754394	0,161625	4,66756	0,0001
ARROWS (RIVAL)	-0,675719	0,128256	-5,26852	0,0000
Attachments (Rival)	-0,773499	0,135867	-5,69305	0,0000

R-square (adjusted for g.l.) = 92.7135 percent

POINTS OBTAINED = 0.026012 *% POSSESSION + 0,702837 * ARC SHOTS + 0,754394 * SHOOTERS - 0,675719 * ARC SHOTS (RIVAL) - 0,773499 * STATEMENTS (RIVAL)

Regression model that relates the average position of the players in the field with the points obtained in the match:

Table 6
ANOVA tables of the regression model associated to the variable position of the equipment and the points obtained by the equipment

Source	*Sum of Squares*	*Gl*	*Squared Mean*	*Reason-F*	*P-value*
Model	147,787	1	147,787	132,07	0,0000
Residue	30,2126	27	1,11899		
Total	178,0	28			

Parameter	*Estimation*	*Error Standard*	*Statistical T*	*P-value*
3/4 DEFENSIVE	1,07874	0,0938664	11,4923	0,0000

R-square (adjusted for g.l.) = 83.0266 percent

POINTS OBTAINED = 1,07874 * 3/4 DEFENSIVE

5. CONCLUSIONS AND DISCUSSION

As preliminary results it is necessary that of the squad of 30 players 9 presented significant correlation with the results of the disputed party, taking into account that the defenses Branislav Ivanovic and John Terry played all the parties, therefore the correlation is presented according to the points Obtained by the team with participation of these two players. In the case of the goalkeepers it is necessary that between Courtois and Cech is 100% of the minutes played, so the expected correlation was obvious, but what is striking is that the coefficient that accompanies Courtois (0.0236978) is less than The coefficient of Cech (0.0337486), although it can be argued that the rivals they faced are different in difficulty, but certainly from the model it is argued that the presence of Cech on the court had a greater positive impact on the result than Courtois.

In the case of the defenses, it is necessary that the players Azpilicueta and F. Luis would be the ideal company of Ivanovic and Terry, although Ivanovic and F. Luis occupied the same position of right side, Ivanovic can play like central defender, therefore the Four players who turned out to be significant could be at the same

time on the court, and together they would explain 85.6% of the variability of the points obtained by Chelsea in the disputed matches, the highest percentage along with the fliers (85.2 %).

In the case of the flyers, players Hazard and Matic were statistically associated with the results achieved by the team, where Hazard is the most technically gifted offensive midfielder and Matic is the defensive midfielder with more minutes played in the team during the 28 days analyzed in the model.

Finally, the most important strikers were Costa, Drogba and Rémy, but they explain only 73.7% of the variability of the points achieved day after day, which suggests the weight of defensive and steering positions In the performance of the team, although followed closely by the goal (the same explained 83.2% of the variability) during the season analyzed where it is possible to point out that the team was crowned champion.

It should be noted that of the 9 players who turned out to be significant in the regression models, 3 left the team in the 2015/2016 season, these were Cech, F. Luis and Drogba, in addition that Rémy has only played a single title and 2 as a substitute, which means that of the 9 players practically 4 are absent, obtaining that of 8 matches played the equipment has obtained 2 victories, 2 ties and 4 losses, which speaks of a yield of 33.3% when in the Season 2014/2015 performance was 76.3%, so you can associate this slump in performance in the possible lack of these players.

In relation to the variables associated with the behavior of the team in the field, the result is in accordance with the logic that governs the development of the game, that the percentage of possession of the ball as a significant variable appears to be something associated with the intention of the team to attack and by In the same way as the number of shots fired at the bow, in the same direction is the number of goalkeeper shortcuts, given that they are avoiding situations that could potentially affect the outcome of the match, in Change the archery shots and saves, from the perspective of the rival also presents statistical significance only that these have a negative effect on the performance of Chelsea, that is to say, the more shots made by the rival team and more attacked by his goalkeeper, Chelsea are more likely to get a bad result.

Of special interest was the result of the average position of the players by sector of the field, in this case the sector of ¾ of defensive field was the only significant one, that is to say as the defensive block of the equipment stays in that position High in the field of play, diminishes the spaces to the rival team and can exert more pressure on the one in search of either that they make a mistake and generate plays of goal or in search of recovering the ball near the rival goal and thus to have majors Scoring options and therefore get better results in the match.

These preliminary results undoubtedly give a deeper understanding of what factors have a greater preponderance in obtaining good results in the game and therefore give the football coach a valuable tool for analyzing the individual and collective performance of their players.

REFERENCES

[1] Castellano, J., & Álvarez, D. (2013). Uso defensivo del espacio de interacción en fútbol. (Defensive use of the interaction space in soccer). RICYDE. Revista Internacional De Ciencias Del Deporte. Doi:10.5232/Ricyde, 9(32), 126-136.

[2] Cavalera. C., et al. (2015) Detección de T-pattern en los partidos de fútbol: Relación entre las acciones del equipo y de ataque. Cuadernos de psicología del deporte. 15(1), 41-50.

[3] Erkmen, N. (2009). Evaluating the heading in professional soccer players by playing positions. Journal of Strength and Conditioning Research, 23(6), 1723-8.

[4] Taskin, H. (2008). Evaluating sprinting ability, density of acceleration, and speed dribbling ability of professional soccer players with respect to their positions. Journal of Strength and Conditioning Research, 22(5), 1481-6.

[5] Casal Sanjurjo, C., Losada López, J. L., & Ardá Suárez, T. (2015). Análisis de los factores de rendimiento de las transiciones ofensivas en el fútbol de alto nivel. Rev. psicol. deport, 103-110.

[6] Severini, T. A. (2014). Analytic Methods in Sports: Using Mathematics and Statistics to Understand Data from Baseball, Football, Basketball, and Other Sports. CRC Press.

[7] Vales Vázquez, Á., Areces Gayo, A., Blanco Pita, H., & Arce Fernández, C. (2015). Perfiles de rendimiento de selecciones ganadoras y perdedoras en el Mundial de fútbol Sudáfrica 2010. Revista de psicología del deporte, 24(1), 0111-118.

[8] Lagos, C., et al. (2010). The Influence of Match Location, the Quality of Opposition and Match Status on Possession in Professional Football. Apuntes Educación Física y Deportes. (102), 78-86.

[9] Frencken, W. G. P. Lemmink, K. A. P. M., & Delleman, N. J. (2010). Soccer-specific accuracy and validity of the local position measurement (LPM) system. Journal of Science and Medicine in Sport, 13(6), 641-5.

[10] Lillestol, J., & Andersson, J. (2011). The Z-poisson distribution with application to the modelling of soccer score probabilities. Statistical Modelling, 11(6), 507-522.

[11] Viloria, A; Parody, A. (2016). Methodology for Obtaining a Predictive Model Academic Performance of Students from First Partial Note and Percentage of Absence. Indian Journal of Science and Technology, Vol 9(46). 1-5.