Pesquisa Operacional (2011) 31(3): 459-465

Printed version ISSN 0101-7438 / Online version ISSN 1678-5142
www.scielo.br/pope

LOGIT MODELS FOR THE PROBABILITY OF WINNING FOOTBALL GAMES

Alessandro Martins Alves, João Carlos Correia Baptista Soares de Mello,
Thiago Graça Ramos and Annibal Parracho Sant'Anna*

Received December 18, 2009 / Accepted December 8, 2010

ABSTRACT. Two ordinal logit models are applied to fit the results of matches in the Brazilian football championship. As explanatory variables are employed measures of previous performance of the teams along all preceding games, along recent games and when playing at home and as a visitor. The results of the models adjustment are employed in simulations performed to forecast the number of points to be earned in the following games and to anticipate the teams' final classification.

Keywords: ordinal logit model, simulation, prediction, football.

1 INTRODUCTION

Organizations, as well as individuals, face an increasing need to reflect on routes to be taken in the future, about what to do or what direction to follow (Roche, 2007). In other words, they are forced to plan. Forecasting the results of football matches has a crucial role for the management of teams in a football league, in planning their posture and deciding on their evolution.

In recent years, efforts have been devoted to forecast results of games for the development of strategies for betting on sports. See, for instance, Craig & Hall (1994) and Lee (1997), who studied the English football, or Stefani & Clarke (1992), for the Australian football.

Scarf & Shi (2008) evaluated the importance of league games in England from 2001 to 2004. Based on such forecasts, the team management would be able to determine the effort to be invested on each game. Depending on the importance of the game for the final rank of the team in the championship, it may be better saving some key players for other matches that the team may be facing. The estimate of the importance of each game is vital also to choose which games televisions stations will show.

Results forecasting is an attempt to predict the future behavior from already observed data. For purposes of forecasting, it is critical to identify which variables or information is relevant and should be included in the predictive models. In football, one such variable is home advantage,

*Corresponding author
Universidade Federal Fluminense – E-mail: annibal.parracho@gmail.com

i.e., the advantage of playing at home. In recent years, in various sports, research has been deepened to verify the existence of home advantage. That such advantage impacts on the outcome of games is supported by several papers, considering not only football, but also baseball, volleyball, American football, among other sports (Courneya & Carron, 1992; Pollard, 2002).

Previous studies, like Pollard (1986), noted that teams playing at home prevail, earning about 60% to 67% of the points disputed in the games played at home. Pollard (2002) examines the reduction of the advantage of playing at home when the club moves from one city to another, what is very frequent in sports where the teams are franchises, or when the team for some reason loses the right to play at home. Nevill *et al.* (1996), studying English and Scottish football leagues for the 1992-1993 season, verified how the presence of the fans impacts on the outcome of the match. Nevill *et al.* (1996) makes a comparison between the divisions of the league, noticing that with a larger public presence, the advantage of the team playing at home was larger. In that paper, two explanations ae provided for that, the weight of the crowd pressure forcing the visitors to make more mistakes and forcing the referee to apply more penalties on the visitor supposed foul plays.

The home advantage may be attributed also to familiarity with the field and stadium, as stated by Courneya & Carron (1992). Pollard (2002, 2006) noticed that the greater awareness of the athlete's home field allows him to take more effective actions along the game.

The change of design of the Brazilian championship to the double round Robin format (all teams playing against all twice), consistent use of the final table of the championship in the selection for further competitions, as well as the relegation of the clubs classified in the worst positions, raised interest in better understanding the chances of each team of not only becoming the champion, but also saving from relegation and being among those better classified (Calôba & Lins, 2006; Sant'Anna, 2008; Sant'Anna *et al.*, 2010). The present work attempts to model the development of 2008 Brazilian Championship aiming at monitoring the performance of each team throughout the championship. Evaluations are performed of the importance on the results of games of home advantage and of the strength of each team. From this, are derived the chances of relegation and of classification for the America Liberators Cup for each club along the championships.

The model applied is the ordinal logit model, also used by Lawall & Sundheim (2007) in modeling the 2002-2003 English Professional League (EPL). The model of Lawall & Sundheim (2007) was effective in anticipating EPL results. That model uses only the advantage of playing at home and the individual strengths of each team as explanatory variables.

Alves *et al.* (2008) applied a similar model to forecast the Brazilian championship. It was noticed then that the model based on home advantage and strength of each team had a deficiency of delaying the capture of changes in the behavior of teams. This delay may be attributed to all previous games being assigned the same predictive importance to the model. The whole history of the team along the championship was given the same importance, while it may be thought that the last games should provide more profitable information. Aiming to capture the "momentum" experienced by each team, a variable is added here to the model with information on the recent performance of the teams, both at home and outside.

The main objective of this study is evaluating the ability of the ordinal logit model fitted at the end of each round to predict the chances of a team being ranked for the America Liberators Cup, *i.e.*, finishing the championship ranked among the first four, and of being relegated, that means, finishing among the four worst ranked. This will be done with the use of two Ordinal Logit models and with the help of simulations based on the models fit.

The next section presents the data analyzed. Section 3 presents the model developed to the present study. Section 4 discusses the results of the application of this model. Section 5 concludes the paper.

2 THE DATA SET

This paper analyses data from the Brazilian championship of the year 2008, which had 20 teams and a total of 380 games. The score for each game is given by: three points for a win, one point for a tie and zero if defeated. For a discussion of this kind of points assignment, see Bloyce & Murphy (2008).

At the end of the competition, the team with the highest total score will be the champion, the four best placed teams ensure a place in next year America Liberators Cup and the four teams with the lowest scores are automatically relegated.

Were considered as playing home those teams thus signaled in CBF's official schedule. It is known that in the cities of Rio de Janeiro and Belo Horizonte, these teams do not benefit from real advantage when facing teams of their hometown because these games are played in a neutral stadium. This may have reduced the estimate of the home advantage variable in the models fitted.

3 THE MODEL PROPOSED

To take into account the three-fold feature of the outcomes of a football games, Lawal (2002) and Lawal and employed an Ordinal Logit model with explanatory variables signaling the effect of playing at home or away. The same model was employed by Alves *et al.* (2008) in an analysis of the Brazilian championship. Here, another explanatory variable, namely, the recent performance of the team, is added to this last model. The model is then given by equations (1) and (2):

$$\ln\left[\frac{P(W_j/X_1\ldots X_{20})}{1-P(W_j/X_1\ldots X_{20})}\right] = \alpha_1 + \sum_{i=1}^{20}\left(\phi_i X_i + \beta_i Z_i\right) \tag{1}$$

$$\ln\left[\frac{1-P(L_j/X_1\ldots X_{20})}{P(L_j/X_1\ldots X_{20})}\right] = \alpha_2 + \sum_{i=1}^{20}\left(\phi_i' X_i + \beta_i' Z_i\right) \tag{2}$$

where 20 is the number of teams in the league,

X_i is given by $X_i = 1$ for $i = j$ and the i-th team playing at home, $X_i = -1$ for $i = j$ and the i-th team playing away and $X_i = 0$ otherwise, for i varying from 1 to 20,

Z_i is given by the percent of points won by the i-th team in the last 4 games played at home if $i = j$ and the j-th team is playing at home, by the percent of points won by the i-th team in the

last 4 games played away if $i = j$ and the j-th team is playing away and $Z_i = 0$, otherwise, for i varying from 1 to 20,

W_i is the indicator of the event of the j-th team winning and L_j, the indicator of the event of the j-th team losing, for j varying from 1 to 20.

Thus, in this model, β_i weights the recent performance of the team. It is assumed that

$$\phi_i' = \phi_i, \quad \forall i = 1, \ldots, 20 \quad \text{and} \tag{3}$$

$$\beta_i' = \beta_i, \quad \forall i = 1, 2, \ldots, 20. \tag{4}$$

Equations (3) and (4) mean that the increase in the odds ratio as the probability of a win is replaced by the probability of a win or a draw does not depend on the strength of the teams on the long run, as well as on its recent strength.

Simulations based on the two models were performed. The first (Model 1) is a simpler version, considering as predictive variables, the home advantage and strength of each team, without considering the recent performance of the team. This is the model successfully employed by Lawal & Sundheim (2007) to model 2002-2003 EPL and Alves (2008) to predict the results of the 2007 Brazilian championship. The results of simulations based on such model are compared to those of the new model (Model 2), which, in addition to the explanatory variables of the previous model, includes the variable evaluating the recent performance of the teams in the competition.

The use of simulation techniques as an alternative to more conventional inference procedures is being used in Statistics for a long time now (Watson & Blacksone, 1989). With the increased processing speed of today computers, this approach has gained increasing importance. Simulations are performed in the present work to derive probabilities of future events from parameter estimates based on the results of the previous games of the championship. At each round, a model is adjusted and the probability of each possible outcome for every future game is calculated.

To each game in the future rounds is assigned a random number between 0 and 1 and, based on the probability derived from the model and the value of such random number, is assigned a result for the game. Based on the results predicted in such way and the results of the games already observed, the final number of points scored by each team in the championship is anticipated. This allows for forecasting which teams will be selected for higher level tournaments and for relegation.

4 ANALISIS OF THE RESULTS

The Home Advantage observed in the 2008 Brazilian championship by round 35, that means, after 350 games played, was 80.57%, *i.e.* in 282 from those 350 games, the team playing at home obtained a win or a draw. When considering only wins of the team at home, this percentage goes to 55.71%, corresponding to 195 of 350 games.

Eight model adjustments were performed, with estimation of models 1 and 2 by the end of rounds 20, 25, 30 and 35. That means, at equal spaces of 5 runs, a new estimation was performed. In all adjustments, the restrictions of equal coefficients for the two equations are met. In order

to forecast the teams selected for ascent and relegation, at each of these four points, 10000 simulations of the results of the following games were performed and the percent of times each team obtained a total of points among the first four and among the last for was calculated.

Table 1 below shows the percentage that each team appeared among the 4 teams with the highest scores after 10000 simulations of the results, considering the likelihood of results derived from the models fit at each round. Were kept in this table only the teams that had a percent of at least 1% of classifications among the first four in at least one of the four moments considered.

Table 1 – % of times each team appears among the first 4.

		Round 20		Round 25		Round 30		Round 35	
	Final punctuation	Mod 1	Mod 2	Mod 1	Mod 2	Mod 1	Mod 2	Mod 1	Mod 2
São Paulo	**75**	**24,6%**	**43,7%**	**38,7%**	**45,0%**	**78,7%**	**79,5%**	**100%**	**100%**
Grêmio	**72**	**99,9%**	**100%**	**98,1%**	**99,0%**	**95,8%**	**96,0%**	**99,9%**	**99,2%**
Cruzeiro	**67**	**84,7%**	**83,6%**	**53,1%**	**48,3%**	**95,2%**	**87,8%**	**97,7%**	**68,6%**
Palmeiras	**65**	**81,3%**	**60,8%**	**92,1%**	**88,6%**	**88,4%**	**90,9%**	**93,2%**	**58,1%**
Flamengo	64	7,6%	7,8%	30,4%	28,2%	38,8%	37,5%	*9,2%*	*74,1%*
Internacional	54	0,0%	0,1%	1,2%	0,6%	0,9%	0,8%	0,0%	0,0%
Botafogo	53	59,3%	27,7%	60,5%	57,8%	0,9%	1,5%	0,0%	0,0%
Coritiba	53	32,0%	14,8%	10,8%	13,9%	0,9%	5,6%	0,0%	0,0%
Goiás	53	0,0%	0,0%	*1,1%*	*0,1%*	0,4%	0,4%	0,0%	0,0%
Sport	52	0,0%	1,5%	*7,7%*	*8,2%*	0,0%	0,0%	0,0%	0,0%
Vitória	52	10,6%	59,3%	6,3%	10,3%	0,0%	0,0%	0,0%	0,0%

By comparing the simulations derived from models 1 and 2, it is observed that model 2 captures earlier than model 1 changes to follow. For instance:

- Until round 10, São Paulo did not give much attention to the Brazilian championship, as had a main objective as a participant of 2008 America Liberators Cup. Thus, by rounds 20 and 25, the models do not place São Paulo among those classified for the next America Liberators Cup. But the percentage given by model 2 is much larger than that given by model 1.

- Flamengo had an excellent performance in the final rounds of the championship (except in the last three); model 1 does not capture the dash, leaving Flamengo always with a very low number of times among the four best ranked, while model 2 by round 35 forecasted its presence among those selected to America Liberators Cup. It is true that Flamengo did not classify, losing the seat to Palmeiras and Cruzeiro, which had by then a lower probability according to model 2. Flamengo stumbled in the last 3 games, with a performance below those that the team had shown. The team lost in round 36 to Cruzeiro, a direct rival in the fight for classification, tied Goiás in round 37 and lost to Atlético-MG in round 38, the final round.

- Like in the preceding year, Botafogo presented a good performance at the start of the championship, but along the championship the team suffered losses and presented a performance inferior to that presented at the beginning; so its chances of finishing among those with the four highest scores were reduced along the time. The biggest drop was on rounds 21 to 25, where the team managed to get only one win, one draw and three defeats. This fall was anticipated by model 2 by the end of round 20.

Table 2 shows percentages of being among the last four by the end of the championship. Table 2 shows that both models showed excellent predictions for the outcome of the championship by round 35.

Table 2 – % of times each team appears among the last 4.

		Round 20		Round 25		Round 30		Round 35	
	Final punctuation	Mod 1	Mod 2	Mod 1	Mod 2	Mod 1	Mod 2	Mod 1	Mod 2
Goiás	53	1,3%	14,7%	0,0%	0,2%	0,0%	0,0%	0,0%	0,0%
Atlético-MG	48	1,2%	0,3%	2,7%	7,2%	0,4%	0,5%	0,0%	0,0%
Atlético-PR	45	39,7%	33,0%	67,1%	40,0%	90,9%	76,5%	6,8%	14,3%
Fluminense	45	82,8%	97,0%	52,7%	60,4%	22,9%	41,1%	0,3%	12,0%
Santos	45	66,4%	63,1%	9,5%	16,3%	0,7%	1,0%	0,0%	0,2%
Náutico	44	50,7%	68,0%	20,1%	34,8%	31,1%	50,6%	1,2%	9,4%
Figueirense	**44**	0,1%	0,0%	6,3%	7,5%	2,0%	2,4%	100,0%	90,2%
Vasco	**40**	27,5%	8,9%	60,4%	54,2%	96,6%	91,0%	91,9%	86,8%
Portuguesa	**38**	33,5%	15,9%	92,9%	87,5%	62,1%	47,8%	99,8%	87,4%
Ipatinga	**35**	96,5%	98,8%	88,3%	91,9%	93,3%	89,1%	100,0%	99,7%

5 CONCLUSIONS

Both models here studied presented good performances when their results were applied to predict the classification of the clubs by the end of the championship. Model 1 presents more stable predictions but is slow in taking into account changes in the behavior of teams. It was verified that considering the recent performance of the team as an explanatory variable improves the model in that sense.

Other variables may be employed to take into account other aspects and details. For instance, the teams strength measured globally in terms of points earned may be divided into offensive and defensive strength, measured respectively in terms of goals scored and goals taken.

Other improvement alternative consists of modeling expected changes. One such change may be derived from the fact that the Brazilian championship is divided into two shifts of 19 rounds and between such shifts there is a gap when several players are drawn by clubs from outside the country. Thus the teams with the best performance along the first half have a great chance of suffering disruptive changes. This information was not directly taken into account in any of the models. It may be explored by using dummies to evaluate whether the differences in the cast of players, or even the exchange of management, can influence the performance of the team.

REFERENCES

[1] ALVES AM, RAMOS TG, SANT'ANNA AP & SOARES DE MELLO JCCB. 2008. Uma proposta de previsão de resultados para o campeonato brasileiro de futebol através do modelo logito. *Anais do SPOLM 2008*. Rio de Janeiro.

[2] BLOYCE D & MURPHY P. 2008. Sports administration on the hoof: the three points for a win "experiment" in English soccer. *Soccer and Society*, **9**: 95–113.

[3] CALÔBA GM & LINS MPE. 2006. Performance assessment of the soccer teams in Brazil using DEA. *Pesquisa Operacional*, **26**: 521–536.

[4] COURNEYA KS & CARRON AV. 1992. The home advantage in sport competitions: a literature review. *Journal of Sport and Exercise Psychology*, **14**: 13–27.

[5] CRAIG LA & HALL AR. 1994. Trying out for the team: Do exhibitions matter? Evidence from the National Football League. *Journal of the American Statistical Association*, **89**: 1091–1099.

[6] LAWAL HB. 2002. Modelling the 1984-1993 American League Baseball Results as dependent categorical data. *Math. Scientist*, **27**: 53–66.

[7] LAWAL HB & SUNDHEIM AR. 2007. Modeling 2002-2003 English Premier League Results. *IMA Sport 2007*, 115–124.

[8] LEE AJ. 1997. Modeling scores in the premier league: Is Manchester United really the best? *Chance*, **10**: 15–19.

[9] NEVILL AM, NEWELL SM & GALE S. 1996. Factors Associated With Home Advantage in English and Scottish Soccer Matches. *Journal of Sports Science*, **14**: 181–186.

[10] POLLARD R. 1986. Home advantage in soccer: a retrospective analysis. *Journal of Sports Sciences*, **4**: 237–248.

[11] POLLARD R. 2002. Evidence of a reduced home advantage when a team moves to a new stadium. *Journal of Sports Sciences*, **20**: 969–973.

[12] POLLARD R. 2006. Worldwide regional variations in home advantage in association football. *Journal of Sports Sciences*, **24**: 231–240.

[13] ROCHE FP. 2007. La planificación estratégica en las organizaciones deportivas, 4th ed. Editorial Paidotribo, Barcelona.

[14] SANT'ANNA AP. 2008. Rough sets analysis with antisymmetric and intransitive attributes: Classification of Brazilian soccer clubs. *Pesquisa Operacional*, **28**: 217–230.

[15] SANT'ANNA AP, BARBOZA EU & SOARES DE MELLO JCCB. 2010. Classification of the teams in the Brazilian Soccer Championship by probabilistic criteria composition. *Soccer and Society*, **11**: 261–276.

[16] SCARFF PA & SHI X. 2008. The importance of a match in a tournament. *Computers & Operations Research*, **35**: 2406–2418.

[17] STEFANI RTE & CLARKE SR. 1992. Predictions and home advantage for Australian Rules football. *Journal of Applied Statistics*, **9**: 251–261.

[18] WATSON HJ & BLACKSTONE JR JH. 1989. Computer simulation, 2nd ed. Wiley & Sons. New York.