ECO 5120 Econometric Theory & Application (2005)

Parametric Modeling of Soccer Goal Time

Ko Chiu Yu[†]

Ng Shou Zhong[‡]

December 2005

Abstract

In recent applied game theory economic research, probability of scoring is expressed as function of strategic and situational variables. Statistical and probabilistic arguments are used to explain the scoring time in statistical journals. This project extends previous studies by combining theoretical results and statistical findings into econometric models. In our project, parametric model adopted is analyzed by survival analysis techniques.

Keywords: Game theory, sports, soccer, survival analysis, Weibull

We would like to thank Professor Lee Myoung-Jae for commenting on an early draft of this paper. Needless to say, any remaining errors are our own.

[†] Chiu Yu, Ko is currently MPhil candidate (Student ID 05168070), Economics Department, CUHK; Lazy@cuhk.edu.hk.

[‡] Shou Zhong, Ng is also currently MPhil candidate (Student ID 05168560), Economics Department, CUHK; s051685@mailserv.cuhk.edu.hk

1. Introduction

In the past two decades, sports matches are studied by economists for empirical testing of predictions using game-theory-based model.[1] At the same time, due to the prevalence of soccer betting, statisticians and mathematicians have tried to model a match in stochastic framework.[2]

Previous empirical analysis by game theorists have suggested that teams' skills, current score, net goal and home field advantage are four significant explanatory variables of the probability of scoring.[3] On the other hand, the empirical models used in statistical studies have shown usual Poisson, negative binomial and mixed passion would be candidates of practical models to predict the number of score in a single football match.[4]

This paper tries combine both viewpoints under one single-framework in survival model and tests the model by fitting real world data. Our analysis yields three main results. First, consistent with previous game theory studies, the team with home field advantage requires less time to make a score. Our finding reinforces previous studies results that surrounding environment has strong influence on players are mainly due to psychological behavior but not strategic behavior.[5] Second, skill of team member is significant in reducing the time to score. However, we found out that the effect of team's skill is not that large as we would have expected, nor coincide with the theories predicted. Third, the explanatory variables net goal shows inverse relationship with the goal time. This is completely different from the predictions derived by game-theory-based models.

[1] Probably, the first few studies are toward strategic choices across games. Like Walker and Wooders [1998] study mini-max hypothersis in tennis games and Ferral and Smith [1999] tested distribution of points in tennis. Later, Palomin, Rigotto, Rustichini [2000] studied soccer focusing strategies across games. .

[2] There are various soccer betting studies in various journals. e.g. Journal of operational Research

[3] See Palomino, F., Rigotti, L. and Rustichini, A. [2000] and reference therein.

[4] See AD Fitt, CJ Howls and M Kabelka, [2005].

[5] Our studies have taken the strategic variable as explanatory variables.

2. Data Sample[6]

We have chosen Barclays English Primer League as our data sample. It is because the data collected is convenient and reliable, as details of matches listed on the official League site and various fan clubs and betting site. [7] We focus on the period from year 2001 to year 2004 totaled 608 matches. The reason for not including current year figure is to test our estimated model by this year data.[8] Due to heterogeneity in the team skills, Liverpool, Chelsea, Manchester United and Arsenal are being chosen to minimize the differential of ability across teams.[9]

The explained variable is scoring goal time which needs special treatments before we could use it. Firstly, it is usually the case that there is no goal in a single match. Then we would treat this censored data at 90 min. If there is one goal in 30 minutes, there would be two observations. One is failure at 30 minutes and the other is censored at 60 minutes. Of course, in the first case censoring and second case censoring is different due to effect of first goal, so we will include the current goal to control this effect.[10]

To explain goal time, we have collect six categories of instruments, namely, match specific, field specific, strategy specific, last matches results, own team specific and opponent team specific. [11]Match specific contains factors invariant throughout a match, for example, climate of the day of match, home field advantage and attendance ratio. Field specific are those factors changing during the match. Current score, current goal (home), current goal (away), current goal conceded (home), current goal conceded (away) are collected to control the effect of changing match environment. Strategic

[6] Dataset are entered by the authors and could be obtained upon request. The data sources are from the following websites: 1.) http://www.premierleague.com. 2.) http://stats.premierleague.com.
[7] Another reason is that previous studies are focused on English Premier League. Then, our founding could be easily compared to other research results.
[8] Since games of the season have not yet ended, we postpone the test under full sample avaible.
[9] These four teams are among the top in the league.
[10] Obviously there is a significant relationship between the scoring time and no. of current goal. In this project, we assume they follow a linear relationship and the explanatory variable "current goal" is added in the model to adjust for this relationship.
[11] See Appendix for detailed description of explanatory variables

specific variables try to control the effect of different strategies adopted. The proxies are last year of fouls accumulated, last year yellow cards/ red cards obtained, last year total goal in the league and last year total goal concealed. The first two are trying to capture the rogue of players and the last two is the offensive-defensive style of the team. Last match effects are being measured by weighted average of previous three game results [12] and current position in the league. The opponent team effect is controlled by last year position of the team in the league and newly promoted team effect. The own team effect[13] is being controlled by the last year position in the League. Since the four teams chosen are remained in League from 2001 to 2004, there is no newly promoted team variable here.

3. Estimated Model

Our final model is based on Weilbull survival model,[14] the basic reasons is underlying hazard rate is time-varying due the simultaneous and interacting forces of strategic rationality and psychological elements. [15]

The final model is shown on Table A. We have found out that only a few explanatory variables are relevant.

[12] The weighting of last match is 3, the match before last match is 2 and 1 for the 2 match before the last match with wining get 3, 0 for draw and -3 for lose.

[13] Own team means the team that the scoring time we are interested in, namely, Manchester United, Arsenal, Liverpool and Chelsea.

[14] Another main reason is that Weibull is a popular choice for duration dependence hazard model. Actually, the estimated model parameter would not change much if we assume other type of distributions. Even if we change the assumption to lognormal, Poisson or exponential would not have any significant impact to estimation result. Therefore, it could be said our estimate is robust to the model selection bias. See Appendix II for details of Weibull distribution

[15] The surrounding environment of match is changing every time of the game and hence every single moment of the game is different from the other moment. The usual independent assumption of time-invariant could not be applied here. This could be further reaffirmed by observing the non-parametric estimated hazard function. See Appendix III for further details.

Table A. Weibull survival Model Estimate

Explanatory Variables	Coefficient	Hazard Ratio	P-value
Match specific			
Home Field Advantage**	0.147	1.159	0.077
Field specific			
Current Score**	0.119	1.126	0.094
Net Goal (Home)***	0.157	1.17	0.00
Net Goal (Away)***	0.185	1.203	0.002
Opponent Ability			
Opponent Last year position***	0.011	1.012	0.00
Previous Game effects			
Weighted previous three results	0.003	1.003	0.237
Strategic Specific			
Last Year Goal ***	0.006	1.005	0.00
Last Year Goal Conceded	-0.009	0.99	0.301
Constant Term	-5.243	0.00	0.764
Number of observation	909		
Log likelihood	-1095		
Chi-Square Test	0.000		

**** significant at 0.005 , **significant at 0.01 , Robust standard error Used*

significant for estimation.

It can be seen from the table that home field advantage is significant in reducing the expected time to score. From our model, the home field advantage would increase the team scoring hazard rate 15% higher, holding other factors consistent. This is consistent with game theory prediction of friendly surrounding environment and also matches with our intuition that home team is more likely to score.

Another important finding is that opponent ability is also crucial to the determination of the goal time. This result looks like consistent with previous studies, however, while direction is correct but magnitude is not. The previous empirical studies reveal the skill differentials raise the probability of scoring by factor of 2.2 to 2.3, ceteris paribus,

however, in our model, the raise of hazard is just by 1%[16], yet p-value is less than 0.0001. One of the reasons of our different conclusion may be due to the fact that given the opponents teams and coach skills, the other teams would adjust their offensive-defensive accordingly. That is to say if the opponent team is empowered with strong striker, the other team would be more likely to adopt more conservative strategy than otherwise.

One of the puzzling results is that the explanatory variables *Net Goal* are significant but completely reverses the direction even the magnitude of prediction is correct. From classical theory, the team is winning, with higher net score, would be more likely to adopt defensive strategy and less incentive to score. That is, winning team would be less likely to score to have favorable tradeoff between offend and defend. However, in our model, the hazard ratio is greater than one, which means winning team is more likely to score! This has contradicted our classical assumptions. No matter own team has home advantage or not, the effect are similar with 20% up in the hazard ratio. It is extremely unpleasant result we come up with. One of the possible reasons behind may be due to sampling error. Given our sample is concentrated on the top teams, winning one score is more likely to win more as the own team player is more passion to strike and shoot and the opponents, particularly the goalkeeper would be depressed. This is true particularly if the net goal is large. Imagine the opponent team is winning three

[16] Opponent ability is being measured by the last year position in the league. Given own teams selected are those among top, the position would just enough to reflect the skills differential. In this way, the higher the differential is, the higher the value of the position will be. Therefore, if skills differential really matters, we would expect the hazard coefficient would be greater than one. As it is the case in our model, we therefore say the theory matches our founding in magnitude. One may quote the scale of measurement may be source of problem. Yet, it could not be the case that even we normalize the scale, we are still unable to account for the difference. This could be further supported by looking at the goal time would not differ much even the skill differential is larger, provided we have removed away the "newly promoted team effect". One evidence to support our claim is the marginal effect is coincide with our estimate. See appendix IV for table of marginal effect.

points. Then it would be very unlikely we would be able to turn around, so it would be more likely to play less hard, if not give up. [17]

4. Maximum likelihood scoring period

From the empirical data, we have found out that period 60 minutes to 80 minutes after a game start is the period with the highest probability of scoring. Hazard rate is 1.5 times during 60 minutes to 80 minutes than other period during the game.[18] Note that hazard rate increase from time zero to highest 60 to 80 minutes but drops sharply from 80 minutes to 90 minutes. [19]

This interesting finding might suggest us that the strikers are having highest performance during 60 minutes to 80 minutes. However, the strikers are poor-performed during the last 10 minutes if there is not any goal during the first 80 minutes. This is an understandable result because if there is no goal in the first 80 minutes, the coach and players would believe that the probability to goal is very slim and spend less effort to attack and more to defend.

5. Extension: Beckham's Effect

As noted, our dataset includes Manchester United Club the period from year 2001 to year 2004. David Beckham served as mid-fielder and team leader in Manchester United during year 2001 to year 2002. He has been later then sold to the Real Madrid since year 2003. Given this information, we could analyze the significance of Beckham contribution.

[17] Here, we are suggesting the relationship between net goal and scoring probability are nonlinear concave function, not monotonic decreasing function. We have tried to fit the model with more dummy and quadratic terms. Yet, all of them are insignificant. Therefore, there is not enough ground to claim the relationship and we refrain ourselves to include this result in our final model.

[18] The result is not confined to the parameter model only. The hazard rate obtained through non-parametric Kaplan-Meier Estimator also reflects the similar results.

[19] See Appendix V for table for estimated hazard ratio.

Assuming the presence of Beckham is time-invariant fixed effect to the hazard ratio, we could allow ourselves to apply dummy variable as proxy to his contribution to the game.[20] The adjusted model[21] has shown that the contribution of Beckham to his club is not statistically significant.

6. Conclusion

Home field advantage, skills differential does reducing the scoring time, though magnitude of skill differential would be smaller than other models predicted. The net goal shows an unexpected sign that theory predicted may be due to the psychological effect of players not modeled in the classical model. During the first 60 minutes the scoring probability is increasing and becomes topped during 60 minutes to 80 minutes but dropping sharply in the last 10 minutes. Applying dummy variable as proxy, Beckham effect[22] is found out to be statistically insignificant.

Further extension can include more mid-stream team data to avoid the problem arise from including only strong team. Moreover, variables like formation arrangement and the effect of pre/post half-time can be added into the regression model. Last but not least, the model can be extended by using panel data across different league in different countries and to compare the difference between 1st goal time and 2nd goal time.

[20] The suitability of using time-invariant fixed effect could be justified on the ground that being the mid-fielder and leader, Beckham contribution to his own team could be said to be fixed during the game. Another reason to use fixed effect is that it is at least natural to assume compare to other normal players, his extra contribution could be treated as fixed effect.

[21] See Appendix VI for model specification

[22] Other player effects could also be found in the similar manner. Clubs may be able to do retrospective valuation of their team members.

7. References

AD Fitt, CJ Howls and M Kabelka, [2005]: " Valuation of soccer spread bets", Journal of the Operational Research Society.

C. S. Lam, [2005]: "Survival Analysis of the timing of goals in soccer games", Hong Kong Economic Journal Monthly, pp.68-pp.71.

Isabelle Brocas, Juan D Carrillo, [2002]: "Do the 'Three-point victory' and 'Golden goal' rules make soccer game more exciting? A theoretical analysis of a simple game", Centre for Economic Policy Research, Discussion paper series, no. 3266.

Myoung-jae Lee, [1996]: "Methods of moments and semiparametric econometrics for limited dependent and variable models", New York: Springer.

Palomino, F., Rigotti, L. and Rustichini, A. [2000]: "Skill, Strategy and Passion: An Empirical Analysis of Soccer", CentER.

William H. Greene, [2003]: "Econometric Analysis", Prentice Hall, Fifth edition.

Wooldridge, Jeffrey M., [2002]: "Econometric analysis of cross section and panel data", Cambridge, Mass.: MIT Press.

Appendix I: Table of Explanatory Variables

Category	Proxy Variable	Remarks
Match Specific	Climate	Whether there is rain, sunny or cloudy and the temperature
	Home field advantage	Whether the team is play as home or away
	Attendance ratio	The percentage of audience of the match
Field specific	Current score	Number of total goal, including both own team and opponent
	Current goal home	Number of goal scored by own team with home field advantage
	Current goal away	Number of goal scored by own team without home field advantage
	Current goal conceded home	Number of goal scored by opponent team with home field advantage
	Current goal conceded away	Number of goal scored by opponent team without home field advantage
Strategic Specific	Last year fouls	Number of fouls obtained by own team in the last year League
	Last year yellow/red cards	Number of year cards obtained by own team in the last year League
	Last year goal	Total number of goal scored by own team in the last year League
	Last year goal conceded	Total number of goal scored by opponent team in the last year League
	Last year shoot	Total number of shoot by own team in the last year League
Last Matches	Previous Three Matches Results	Weight average of Last three matches result
	Current Position in the league	Current position in league of own team before the match
Opponent Team	Last year position	Last year position in the League
	Newly promoted team effect	Whether the team is newly promoted
Own Team	Last year position	Last year position in the League

Appendix II: Weibull Distribution

Hazard function:

$$\lambda(t) = \gamma\alpha t^{\alpha-1}$$

Taking Logarithm of both sides yields:

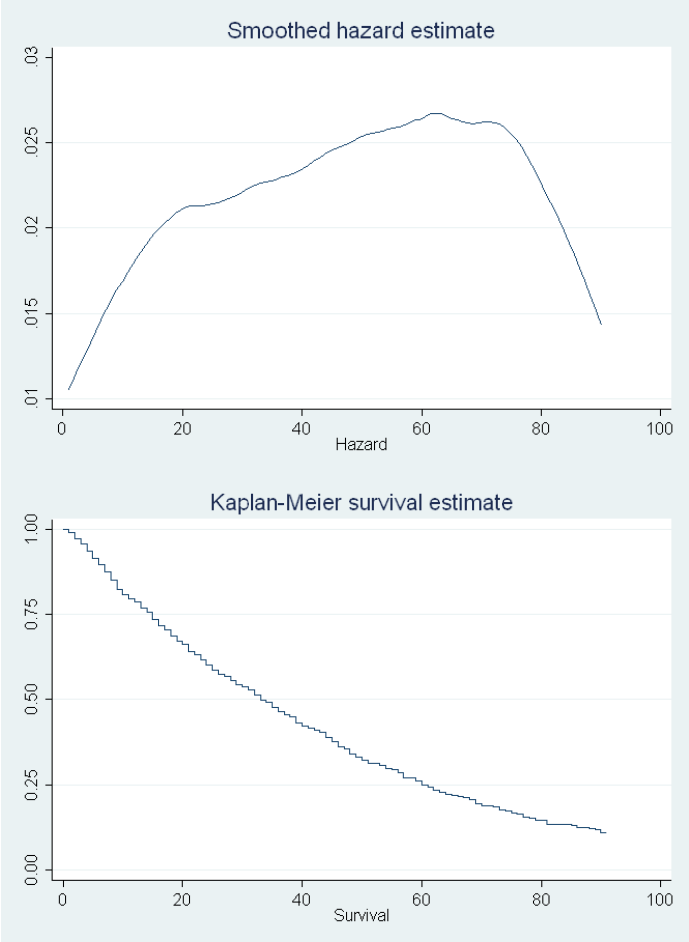
$$\ln\lambda(t) = \ln(\gamma\alpha) + (\alpha - 1)\ln t$$

where $\gamma = \exp(x'\beta)$ and x is a vector of independent variables

Please refer to standard textbook for further details.

Appendix III: Non-parametric estimated model

Using Kaplan-Meier Survival Estimation, we have seen that the hazard function is non constant. The smoothed hazard estimate graph has shown the non-linear relationship between hazard rate and the time. This has support the usage of Weibull distribution in our final model.



Appendix IV: Marginal Effect

Variable	Marginal effect
Opponent ability	-0.305
Home effect	-3.816
Current score	-3.062
Net goal (Home)	-4.045
Net goal (Away)	-4.763
Last year goal	-0.153
Last year goal conceded	0.244
Previous game result	-0.092

Appendix V: Estimated Hazard function across time

The following table shows the hazard ratio estimated from the Weibull survival function across time.

Time interval		Beg.	Cum.	
From	**To**	**Total**	**Failure**	**Hazard**
0	10	912	0.1768	0.0194
10	20	686	0.3281	0.0202
20	30	520	0.4572	0.0213
30	40	383	0.5683	0.0228
40	50	275	0.6668	0.0258
50	60	191	0.7364	0.0233
60	70	135	0.8043	0.0296
70	80	88	0.8523	0.028
80	90	55	0.8789	0.0198

Appendix VI: Model to test marginal contribution of Beckham

The following (Weibull) model is used to test the Beckham's effect:

$$\ln \lambda(t) = \ln \alpha + \beta_1 x_1 + ... + \beta_{Beckham'seffect} x_{beckham'seffect} + (\alpha - 1) \ln t$$

Where $x_{beckham'seffect} = \begin{matrix} 1 \\ 0 \end{matrix}$ if the year $\begin{matrix} 01/02 - 02/03 \\ 03/04 - 04/05 \end{matrix}$

The estimated result is:

Explanatory Variables	**Coefficient**	**P-value**
Match specific		
Home Field Advantage	0.092	0.487
Field specific		
Current Score**	0.174	0.00
Net Goal (Home)***	0.166	0.05
Net Goal (Away)***	0.101	0.173
Opponent Ability		
Opponent Last year position***	0.080	0.420
Previous Game effects		
Weighted previous three results	0.060	0.147
Strategic Specific		
Last Year Goal ***	0.010	0.567
Last Year Goal Conceded	-0.0133	0.368
Constant Term	-5.521	0.00
Beckham's effect	0.108	0.657