

Predicting Soccer League Games using Multinomial Logistic Models

STAT 525 Course Project, Fall 08
Fang Liu and Zheng Zhang
Purdue University

1 Introduction

In recent years, applying statistical methods for analyzing sports data has received much attention. It has many applications such as providing guidance for sports and assisting betting. Guided by several previous papers on this topic, we look into modeling and predicting soccer games. As the most popular sport in the world, soccer is played in many countries in the form of leagues. Their data is readily available on the Internet, facilitating our analysis.

A lot of papers have been published presenting statistical methods for sports data, covering a great variety of sports including soccer. Various models are used, *e.g.*, Poisson models [2, 3]. We focus on soccer played in the league format. Our goal is to develop a model to predict the outcomes of games that are not yet played in a league, based on the model training using games already played. While guided by previous work, we follow a different approach – we use multinomial logistic regression models.

We summarize the highlights of our paper as follows,

- We propose and discuss several multinomial logistic models for modeling and predicting games in a soccer league.
- We show how to transform the models to forms that are friendly to statistical software.
- We evaluate these models by applying model selection on the data of 2007-2008 season English Premier League. The selected best model can accurately predict the outcome of a future game with around 0.60 probability.
- We perform computer simulation on the remaining part of the league and predict the final standing.

Before going into technical details, we introduce some background on soccer league. A soccer league consists of a number of teams, who play against *every* other team twice in one season, once on the home ground and once on the away ground. In the home team's point of view, there are three outcomes of a single soccer game, namely win, draw and loss, in which the home team earns 3, 1 and 0 points and the away team earns 0, 1 and 3 points respectively. Teams are ranked according to the total earned points. Soccer is also played in other formats, such as knockout tournament and multi-stage tournaments.

The rest of the paper is organized as follows. In Section 2 we present our models. In Section 3, we apply our models to soccer data. Section 4 concludes the paper.

2 Modeling Soccer Games

In this section, we propose and discuss several models.

2.1 Conceptual Model of Soccer Games

A soccer game can be modeled as the comparison of the various attributes of two opponents. These attributes can be the attack and defense ability, which can further be broken down into passing, dribbling, shooting and tackling ability of each player, team work and the coach's strategy. To make it simple, we abstract a team i into a single attribute, called *strength* s_i . The outcome of a game G_{ij} between home team i and away team j is related to the strength differences, *i.e.*, $s_i - s_j$. It is known that in soccer games, the home team has an advantage. To account for this advantage, we assume a boost h to the home team's strength. Thus, we model a game as a regression in which the G_{ij} is the response variable related to $h + s_i - s_j$. The exact relationship depends on how model is formulated (implemented). It will be discussed later in Section 2.3.

Our next question is how to obtain s_i and h . One strategy is to assume certain external ranking as the team strength s_i . Another strategy is to assume the strength as unknown parameters and let the model fitting infer them. The latter strategy is more attractive if there is enough data to infer these parameters, and otherwise only the former strategy is feasible. For a single-elimination tournament involving N teams, the number of games is typically $N - 1$. Considering each game as one observation, $N - 1$ games cannot infer N unknown parameters. Therefore, previous work on modeling single-elimination tournaments usually adopt the former strategy, *i.e.*, take the strength from external rankings. For the league format which we focus on, the number of games is typically $N(N - 1)$, which is enough to support inferring N unknown parameters. Similar is h . Therefore, we adopt the latter strategy, *i.e.*, assume s_i and h to be unknown.

2.2 Assumptions

Our modeling is based on the following two assumptions: (1) The strength of a team is fixed throughout the season and regardless of its opponents. (2) Games are conditionally independent given (1). We discuss the appropriateness of the two assumptions. Firstly, assumption (1) is largely valid because the squad of a team is quite stable throughout the season, and thus the team strength is largely stable. However some factors may contribute the fluctuation of the team strength, *e.g.*, morale, fatigue, injury of key players, joining of new player in the middle of the season. One of the limitations of our model is that it does not consider these factors. Secondly, assumption (2) is valid. Given assumption (1) holds, the randomness of a soccer game is due to factors such as weather and referee's fairness. The weather factor is independent across multiple games, because different games are held at different times and locations. The fairness of referee factor is also independent across multiple games, because referees are assigned randomly.

2.3 Nominal Models

As stated in Section 2.1, we model the outcome of a game as a response of team strength difference. We name all the N teams by integer from 1 to N . As above, we denote the outcome of a match between home team i and away team j by G_{ij} . Such a naming scheme is not ambiguous, because team i meet team j only once at i 's home ground. $G_{ij} = 0, 1, 3$ corresponds the home team loss, draw, win. We use $\pi_k(i, j)$ to denote the probability of G_{ij} being k , *i.e.*,

$$\pi_k(i, j) = P(g_{ij} = k), \quad k = 0, 1, 3; \quad \sum_{k=0,1,3} \pi_k(i, j) = 1$$

The team strength difference is quantitative and the game outcome takes on more than two discrete values, which suggests a multinomial logistic model. We now treat the game outcome as nominal. The case that the outcome is treated as ordinal will be discussed in Section 2.5. Our model is formulated as follows,

Model 1 (*Reduced Nominal Game Model*) Each team i has two strength parameters $s_{i,k}$, where $k = 1, 3$. Parameter h_k reflects the home team advantage.

$$\log \frac{\pi_k(i, j)}{\pi_0(i, j)} = s_{i,k} - s_{j,k} + h_k, \quad k = 1, 3 \quad (1)$$

G_{ij} is the response variable, $s_{i,k}$ is a categorical parameter and h_k is also a parameter. To avoid overparameterization, we let $s_{N,k} = 0$, where $k = 1, 3$.

The strength of a team is split into two parameters. The parameter $s_{i,k}$ is interpreted as team i 's strength towards achieving outcome k against outcome 0. Such as strength is relative to that of the last team N . The parameter h_k is interpreted as the home team's chance of achieving outcome k against outcome 0, if home team and away team have the same strength. $h_3 > 0$ reflects that home team has an advantage.

Note that the implication of the splitting strength into two parameters is that defining a total order on team strengths is now difficult. Given two teams i and j , the condition $s_{i,3} > s_{j,3}$ does not ensure that team i has better chance than team j of winning a third team. Whether team i has better chance depends on the strength of the third team. Thus it is not straightforward to define which team is better.

Model 1 assumes that all the home teams have the same advantage. We can also assume that each team i has a different home advantage $(sh)_{i,k}$, which leads to the following full model.

Model 2 (*Full Nominal Game Model*) Each team i has two strength parameters $s_{i,k}$, where $k = 1, 3$. $(sh)_{i,k}$ reflects the home team advantage of team i .

$$\log \frac{\pi_k(i, j)}{\pi_0(i, j)} = s_{i,k} - s_{j,k} + (sh)_{i,k}, \quad k = 1, 3 \quad (2)$$

G_{ij} is the response variable, and $s_{i,k}$ and $(sh)_{i,k}$ are categorical parameters. To avoid overparameterization, we let $s_{N,k} = 0$, where $k = 1, 3$.

We do not consider the interaction between two team strengths, because we assume that a team always has the same strength regardless of its opponents. Otherwise we have too many parameters and the data in a single season cannot support them.

To predict the outcome of a game, we use one of our models to calculate the probabilities of each outcome, namely win, draw and loss, and pick the most probable outcome.

2.4 Computational Methods

We describe the computational methods to fit the Model 1 and Model 2 using statistical software. Neither model is straightforward to express in the language of statistical software. The challenge is caused by the $s_{i,k} - s_{j,k}$ part, where i and j are factor levels belonging to the same factor space. In contrast, typically we encounter $a_i + b_j$ in a two-way ANOVA model, where i and j belong to two different factor spaces, and statement like `class A, B` would express the model in SAS. In order to make statistical software understand our models, we need to reformulate them.

Model 2 is reformulated as,

$$\log \frac{\pi_r(i, j)}{\pi_0(i, j)} = t_{i,k} - s_{j,k}, \quad k = 1, 3 \quad (3)$$

where $t_{i,k} = s_{i,k} + (sh)_{i,r}$ merges two parameters. $t_{i,k}$ and $s_{i,k}$ represent the home play strength and away play strength of team i respectively. This model does not enforce any relationship between $t_{i,k}$ and $s_{i,k}$. Thus, in Equation 3, i belongs to the home team space and j belongs to the away team space. Those two are totally different spaces. Thus the model can be handled by statement like `class A, B` in SAS.

The same trick does not apply to Model 1, because otherwise the resulted model enforces a relationship between $t_{i,k}$ and $s_{i,k}$, i.e., $t_{i,k} - s_{i,k} = h_k$ for all i , which is not present in the original form. We work around this problem by using indicator variables.

For a game between home team i and away team j , we use $N - 1$ indicator variables x_1, \dots, x_{N-1} , where

$$x_r = \begin{cases} 1 & \text{if } r = i \text{ and } i \neq N \\ -1 & \text{if } r = j \text{ and } j \neq N \\ 0 & \text{otherwise} \end{cases} \quad r = 1, 2, \dots, N - 1$$

Then we reformulate the model as the following form, which is a canonical multinomial logistic regression,

$$\log \frac{\pi_r(i, j)}{\pi_0(i, j)} = \sum_{r=1}^{N-1} s_{r,k} x_r + h_k, \quad k = 1, 3 \quad (4)$$

Note that $s_{N,k}$ is not in the equation since the original model assumes $s_{N,k} = 0$.

2.5 Discussions: Ordinal vs. Nominal Models

Model 1 and Model 2 treat the response variable as nominal. The response variable can be also treated as ordinal, because there is an order in win, draw and loss corresponding to the strength difference. Accordingly, we have both reduced and full models.

Model 3 (*Reduced Ordinal Game Model*) Each team i has a strength parameter s_i . h_k reflects the home team advantage, where $k = 1, 3$.

$$\log \frac{\pi_k(i, j)}{\pi_0(i, j)} = s_i - s_j + h_k, \quad k = 1, 3 \quad (5)$$

G_{ij} is the response variable, s_i is a categorical parameter and h_k is also a parameter. To avoid overparameterization, we let $s_N = 0$.

Model 4 (*Full Ordinal Game Model*) Each team i has a strength parameter s_i . $(sh)_{i,k}$ reflects the home team advantage of team i , where $k = 1, 3$.

$$\log \frac{\pi_k(i, j)}{\pi_0(i, j)} = s_i - s_j + (sh)_{i,k}, \quad k = 1, 3 \quad (6)$$

G_{ij} is the response variable, and s_i and $(sh)_{i,k}$ are categorical parameters. To avoid overparameterization, we let $s_N = 0$.

However, we argue that neither nominal nor ordinal models are ideal for implementing the conceptual model in Section 2.1. We state the ideal model as follows.

Model 5 (*Reduced Ideal Game Model*) Each team i has a strength parameter s_i and h reflects the home team advantage.

$$\log \frac{\pi_3(i, j)}{\pi_1(i, j)} = s_i - s_j + h \quad (7)$$

$$\log \frac{\pi_0(i, j)}{\pi_1(i, j)} = s_j - s_i - h \quad (8)$$

G_{ij} is the response variable, s_i is a categorical parameter and h is also a parameter. To avoid overparameterization, we let $s_N = 0$.

The above model is built on the following principle – the logit ratio of win to draw is the difference of (adjusted) strengths of two teams. Note this principle does not state anything about loss. However, applying this principle on the home team side, we obtain Equation 7. Applying this principle on the away team side, we obtain Equation 8. Note that away team win is home team loss.

Unfortunately, such an ideal model does not conform to standard formulation of logistic models, and thus it requires a significant extra effort to solve it. Compared to the ideal model, the ordinal models are restrictive, and nominal models are relaxed. Given that ideal model is not available, it is not clear which of the other four models is the best. We will answer this question using model selection procedure in Section 3.3.

3 Applying the Models

In this section, we apply our models to our data – 2007-2008 season English Premier League. We perform model selection and select Model 1 as the best model. Then we generate and analyze the residual plots. Finally using the selected model, we perform repeated computer simulation of the unfinished 8 rounds of the league, and predict which team is most likely to win the champion.

Model	D.F.	Deviance	Goodness of fit test		AIC	Prediction Accu. (%)	
			Deviance	Pearson		Training	Validation
Reduced Ordinal (3)	425	383.8	0.3996	0.0995	425.8	45.7%	43.2%
Full Ordinal (4)	406	359.4	0.9535	0.1659	439.3	64.1%	55.4%
Reduced Nominal (1)	406	357.0	0.9889	0.0594	437.0	65.5%	59.5%
Full Nominal (2)	368	309.4	0.9882	0.0679	465.4	68.6%	48.6%

Table 1: Model selection details.

3.1 Data Set

We have collected data of 2007-2008 season English Premier League from website SoccerSTATS.com [1]. The website embeds the data in html web pages. We wrote a computer script to extract the data by crawling and parsing these pages. Our data include the home and away teams, number of goals scored and points earned by each team in each game. The league has 20 teams, $2*(20-1) = 38$ rounds, and $20*(20-1) = 380$ games. The league has played 32 rounds, 297 games¹, and will finish on May 2008. Among the 297 games, home teams have 138 wins, 78 draws and 81 losses, clearly indicating a home team advantage. Manchester United, Arsenal and Chelsea rank top three, and Derby County ranks bottom.

¹There is one incomplete round and 3 games in it are postponed.

3.2 Statistical Tools

We process the data using both R and SAS. We use SAS procedure `proc logistic` with option `link=logit` to fit the ordinal models, and option `link=glogit` to fit the nominal models. SAS produces deviance, goodness of fit test, AIC and residual plots. To evaluate the prediction accuracy, we use R to fit these models again. We use the `multinom` procedure provided by `nnet` library to fit nominal models, and `polr` procedure provided by `MASS` library to fit ordinal models. Finally, R is used to perform computer simulation. As discussed in Section 2.4, transformed forms of the original models are input into R and SAS. Source codes are available at <http://www.ece.purdue.edu/~zhang97/stat525prj/>.

3.3 Model Selection

Besides AIC, we also consider the prediction ability as a selection criterion, since our primary interest is to predict. We randomly split the data into a training set and a validation set. 3/4 of the data goes into the training set, and the rest 1/4 goes into the validation set. Prediction accuracy is defined as the percentage of games that are predicted accurately by the model.

Table 1 shows model selection details. We have four observations. First, all four models produce good fit, as indicated by Deviance and Pearson goodness of fit tests. Second, as models become more complicated, the fit on the training data is better, as indicated by lower deviance and higher prediction accuracy on the training data. Third, AIC criterion suggests the reduced ordinal model as the best model and prediction accuracy on the validation data suggests the reduced nominal as the best model. Since our goal is to predict accurately, we prefer the prediction accuracy criterion and thereby select the reduced nominal model. Finally, the selected model produces around 60% prediction accuracy, much better than 30% by random guessing.

3.4 Model Diagnosis

Next we perform diagnosis on the selected reduced nominal model. We fit the model on the entire data set and draw diagnostic residual plots. We separate the deviance residuals to two data sets where the nominal response is considered as two binary responses (result 3 vs 0 and result 1 vs 0) and the two sets of deviance residuals are diagnosed separately. Figure 1(a) and Figure 1(c) show the residuals versus predicted probabilities with lowess smooth. The lowess smooth of the plot is approximately a horizontal line with zero intercept that suggests a goodness of the model fit. Figure 1(b) and Figure 1(d) are the half-normal probability of the residuals. Again, the plot reveals that the linear part of the logistic regression is adequate. In summary, the diagnostic residual plots suggest the adequacy of the model fit.

Another view of the residuals is shown by Table 2. Prediction is performed on all the played games and the predicted outcomes are summarized on per team basis. As in Section 3.3, the outcome of a game is predicted as the most probable result among win, draw and loss as estimated by the model. We see that our model consistently overestimates the number of draws and underestimates the number of wins and losses over almost all teams. The most striking gap is seen on Manchester United, where our model cannot explain why this team has so few draws. This misprediction happens even if in our model design we have explicitly specified separate strength parameters for achieving draws and achieving wins. Not surprisingly, this is because a soccer game is more complicated than simply the subtraction of two team strengths. Perhaps defining a different criterion of mapping the win, draw and loss probabilities to a game outcome, *e.g.*, using cutoffs that bias the draw probability could be a remedy our model.

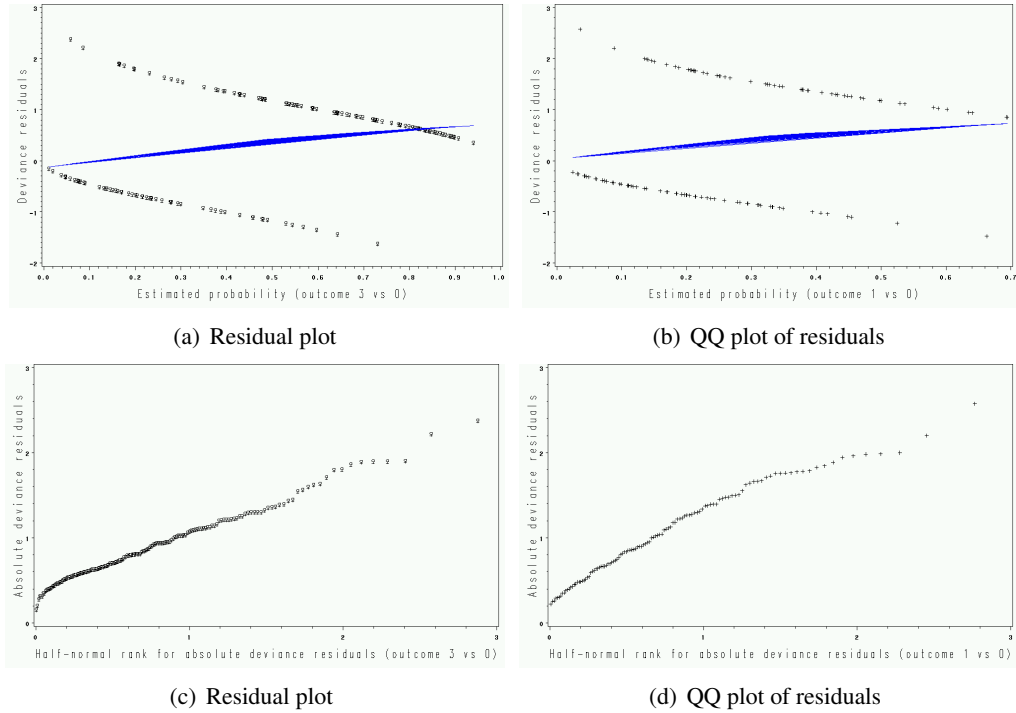


Figure 1: Residual plots of the reduced nominal model.

3.5 Analyzing the Model Parameters

Table 2 also shows the estimates of two strength parameters $s_{i,1}$ and $s_{i,3}$ of each team. Generally, teams that rank higher in the current stage of the league have higher $s_{i,1}$ and $s_{i,3}$. As discussed in Section 2.3, the implication of splitting team strength into two parameters is that there exists no total order of teams in terms of strength. Thus how to directly compare two teams is not straightforward. We look at three teams Manchester United, Arsenal and Chelsea for an example. Chelsea has the largest strength on win, but also has the largest strength on draw. This implies that when playing against the same opponent, Chelsea does not necessarily has higher probability of winning than the other two teams, or higher probability of drawing. Whether Chelsea has higher probability depends on who the opponent is.

Table 3 shows the estimation of parameter h_1 and h_3 . As discussed in Section 2.3, $h_3 > 0$ implies that home team has higher probability of winning than losing when opposing an equal strength away team. Table 3 shows that $h_3 > 0$ is statistically significant, suggesting the existence of home advantage. Note that h_1 alone does not indicate home advantage.

3.6 Predicting the Final Standing

We use computer simulation based on our model to predict the final standing of the league. This involves predicting the outcomes of 83 future games in the remaining 8 rounds. Using the trained model, we estimate the probabilities of outcomes of each future game (π_0, π_1, π_3) . We perform 100 simulation passes on the unfinished rounds. Each simulation pass is performed as follows. To simulate a game, the game outcome is randomly drawn from $\{0, 1, 3\}$ using (π_0, π_1, π_3) as weights. After the simulations of all future matches are done, the points of each team are accumulated to the current earned points, and final standing is then

Teams	Observations							Parameter est.		Predictions		
	Games	Pts	Win	Draw	Loss	Scored	Conceded	$s_{i,1}$	$s_{i,3}$	Win	Draw	Loss
Manchester United	29	67	21	4	4	59	15	1.582	3.028	14	14	1
Arsenal	30	67	19	10	1	58	22	2.303	3.632	16	14	0
Chelsea	29	64	19	7	3	49	18	3.184	3.666	13	13	3
Liverpool	30	59	16	11	3	55	21	1.680	2.387	13	14	3
Everton	30	56	17	5	8	47	25	0.806	1.948	12	13	5
Portsmouth	30	50	14	8	8	44	31	2.750	2.259	5	18	7
Aston Villa	30	49	13	10	7	52	39	-0.351	1.347	12	17	1
Manchester City	30	48	13	9	8	36	34	0.797	1.619	12	12	6
Blackburn	30	46	12	10	8	39	37	2.040	1.986	11	13	6
West Ham United	30	43	12	7	11	33	36	0.749	0.812	7	15	8
Tottenham	29	35	9	8	12	54	47	-0.130	0.123	7	14	8
Wigan Athletic	30	31	8	7	15	27	42	0.000	0.000	6	13	11
Middlesbrough	30	31	7	10	13	27	44	-0.383	0.213	7	12	11
Reading	30	28	8	4	18	35	57	-0.248	0.055	6	12	12
Newcastle	29	28	7	7	15	30	56	-0.067	-0.283	3	13	13
Sunderland	30	27	7	6	17	26	48	-0.373	-0.461	7	9	14
Birmingham City	29	26	6	8	15	33	45	0.361	-0.027	3	12	14
Bolton	29	25	6	7	16	28	43	-0.412	-0.683	3	13	13
Fulham	30	23	4	11	15	27	49	-0.741	-0.442	6	12	12
Derby County	30	10	1	7	22	14	64	-1.042	-1.890	0	15	15

Table 2: Team standing at the current stage of the league, parameter estimation and prediction.

Param.	DF	Est.	Std. Err.	Wald χ^2	$Pr > \chi^2$
h_1	1	0.5601	0.2847	3.8697	0.0492
h_3	1	1.1679	0.2633	19.6790	<.0001

Table 3: Estimation of parameter h_1 and h_3 .

derived. Thus, each simulation pass produces a final standing. Repeating the simulation pass 100 times, we obtain the distribution of final standing as shown in Table 4.

We can see who will win the champion in Table 4. Manchester United has 0.56 probability of winning the champion while Arsenal has 0.24 probability, and Chelsea has 0.20 probability. The prediction that Manchester United has better chances than the other teams is consistent with the current standing.

4 Conclusion

In this paper, we proposed several multinomial logistic regression models to predict the soccer league games. We evaluated and selected our models on data from a soccer league. Overall the selected best model is a good fit, according to the goodness-of-fits statistics and residual plots. More importantly, the model is able to predict the future game with 59.5% accuracy, which is much higher than random guess (33.3%).

Prob. (%)	Standing							
	1	2	3	4	5	6-10	11-15	16-20
Manchester United	56	24	20	0	0	0	0	0
Arsenal	24	41	34	1	0	0	0	0
Chelsea	20	35	45	0	0	0	0	0
Liverpool	0	0	1	72	21	6	0	0
Everton	0	0	0	24	65	11	0	0
Portsmouth	0	0	0	2	4	94	0	0
Aston Villa	0	0	0	1	10	89	0	0
Blackburn	0	0	0	0	0	100	0	0
Manchester City	0	0	0	0	0	100	0	0
West Ham United	0	0	0	0	0	97	3	0
Tottenham	0	0	0	0	0	3	97	0
Middlesbrough	0	0	0	0	0	0	94	6
Reading	0	0	0	0	0	0	83	17
Wigan Athletic	0	0	0	0	0	0	81	19
Newcastle	0	0	0	0	0	0	66	34
Birmingham City	0	0	0	0	0	0	39	61
Sunderland	0	0	0	0	0	0	28	72
Fulham	0	0	0	0	0	0	5	95
Bolton	0	0	0	0	0	0	4	96
Derby County	0	0	0	0	0	0	0	100

Table 4: The distribution of final standing generated by 100 simulations of the unfinished matches of the league.

References

- [1] SoccerSTAT. <http://soccerstat.com>.
- [2] D. Karlis and I. Ntzoufras I. Statistical Modelling for Soccer Games. In *Proceedings of Hellenic European Conference on Computer Mathematics and its Application*, 1998.
- [3] D. Dyte and S. R. Clarke. A Ratings Based Poisson Model for World Cup Soccer Simulation. *Journal of the Operational Research Society*, 51(8), Aug. 2000.