Science in China Series A: Mathematics
Jun., 2009, Vol. 52, No. 6, 1139–1156
www.scichina.com math.scichina.com
www.springerlink.com

© 2009 SCIENCE IN CHINA PRESS
Springer

An analysis of Chinese Super League partial results

Dedicated to the memories of Priscilla Yard Silber (1910–1994) *and Austin Carlyle Brillinger* (1911–1938)*, both born in Sichuan Province*

BRILLINGER David R

Department of Statistics, University of California, Berkeley, CA 94720, USA
(email: brill@stat.berkeley.edu)

Abstract Some of the history of soccer/world football in China is presented. Then consideration turns to the 2008 Chinese Super League. It has 16 teams. The results from the first half of the season, i.e. 15 rounds, are studied. The response of interest for a specific game is whether the home team won, tied or lost, who the home team was, and who the opponent was. The response is ordinal-valued. A generalized linear model is fit and then, given the remaining fixtures, used to predict the final standings of the season. Other explanatories, such as round number, are considered for inclusion in the model. Simulation is employed to estimate probabilities of interest.

Keywords: China, forecasting, ordinal data, simulation, soccer, Super League, world football, 2008

MSC(2000): 62M10, 62-07, 62J12, 91A50

1 Introduction

Sports and statistics are natural companions. There have been analytical investigations with surprising results, e.g. [1] shows that the problem of determining whether a given team still has a chance of being champion is NP-hard. There are journals, e.g. *Journal of Quantitative Analysis in Sports*, books, e.g. [2], reviews, e.g. [3], and there are sections and committees of societies e.g. the American Statistical Association, the Royal Statistical Society, and the International Statistical Institute. Mosteller records the question, "What is the chance that the better team in the series wins?" and notes that "Some people did not understand the concept that there might be a 'best' or 'better' team possibly different from the winner." Statistical analysis can address such issues.

In analytic studies of sports data the techniques of stochastic processes and time series analysis are often employed. At the same time ordinal data, where the sample space has an order relation, is common in practice. This paper considers all three of these variants. The basic datum may be viewed as a results table or 8 by 5 matrix with supplimentary information available on the games remaining to be played.

The paper continues and extends similar analyses for hockey and other football/soccer leagues

Received October 1, 2008; accepted January 5, 2009
DOI: 10.1007/s11425-009-0065-z
This work was supported by US National Science Foundation (Grant No. DMS-0707157)

Citation: Brillinger D R. An analysis of Chinese Super League partial results (with discussion). Sci China Ser A, 2009, 52(6): 1139–1156, DOI: 10.1007/s11425-009-0065-z

working with the basic result of who won, tied or lost (W-T-L) a given game during the so far completed part of a season, see [4–6]. Specifically the paper presents an analysis of the results for the first 15 rounds and the 16 teams of the 2008 Chinese Super League season. The teams' outcomes are viewed as ordinal-valued in accordance with the observed results. There are arguments for this. For example standings and progression to other tournaments are based on these values. Further this approach is directly comparative. Lastly it obviates the need for modelling an increasing scoring rate if one is present. Stochastic models are constructed. The stochasticity is meant to handle things like weather, referee's errors, yellow and red cards, injuries. Explanatories included in the analyses are presented: home team, visiting team, round number, and the previous game results.

Various interesting questions arise. Foremost amongst these is, can one reasonably forecast the remaining game results? Does a team's result in the preceding game matter? Simulations will be employed to study some of these questions. The flexibility of that method is a strong advantage.

For a given game in setting down likelihood functions, a trinomial model with the probabilities consistent with ordinality, is employed for the results of the home teams with the different game results in the same round assumed conditionally independent given the past. The estimates studied are maximum likelihood.

Estimates are computed of the probability of each of the teams being in the top four at the finish of the season. These teams qualify for the Asian Federation Champions League the following season. Further the ultimate champion is eligible for the Asian Champions Cup.

By means of simulation a forecast is derived of the final points that each team will have. One finds the leader, ShaanXi, ahead with an estimated chance of becoming the champion of 24007/25000. The model may be employed to estimate the chances of other interesting events, for example the distribution of the number of ties in a specified round knowing which teams are playing.

The specific structure of the paper is the following: Introduction, Football in China, Some previous work on football/soccer statistics, Ordinal-valued variables, Ordinal-valued time series, Results, Uses and applications, Model fit, Including time, Discussion and summary, Appendix. The Appendix includes a discussion of latent variable and cutpoint models and some details of the computations.

2 Football in China

2.1 History

A case can be made that football was invented in China as long ago as 5000 B.C. In fact football in China appears to have two separate histories, the ancient and the modern. One is the story of *tsu chu*, see [7]. This is an ancient Chinese ritual/game that has been argued as the world's first football game. It was played during the reign of Emperor Huang-Ti (ca. 2500 B.C.). The word "tsu" translates as kicking a ball with one's feet, while "chu" refers to a stuffed ball made of animal skin. Other elements, reconstructed from Han Dynasty sources include goal posts with a net and disallowing the use of hands. The modern history is one of soccer's entering China via European colonists in the mid-1800s. To complete the cycle, in recent years a number of Chinese players have gone on to compete in major European Leagues.

The soccer encyclopedia entry “China, Peoples Republic”[8] is one reference to the history. More detail may be found in *Football in China*[9], which remarks that “Football has been one of the most well supported sports in China ever since it was introduced in the early 1900s”. Also there are three other Wikipedia articles and a popular book[10−12], listed at the end of this paper.

2.2 The Super League

The China Super League was founded in 2004. Its season runs February-March to November-December. There are 16 teams in 2008. The names of the teams are listed in Table 1 and their locations shown in Figure 1. (The order of the teams in Table 1 is that of 20 July when the present analyses commenced.) The teams are located in the southeastern part of the country, see Figure 1. The 2007 champion was Changchun. In Round 15 they are in ninth position.

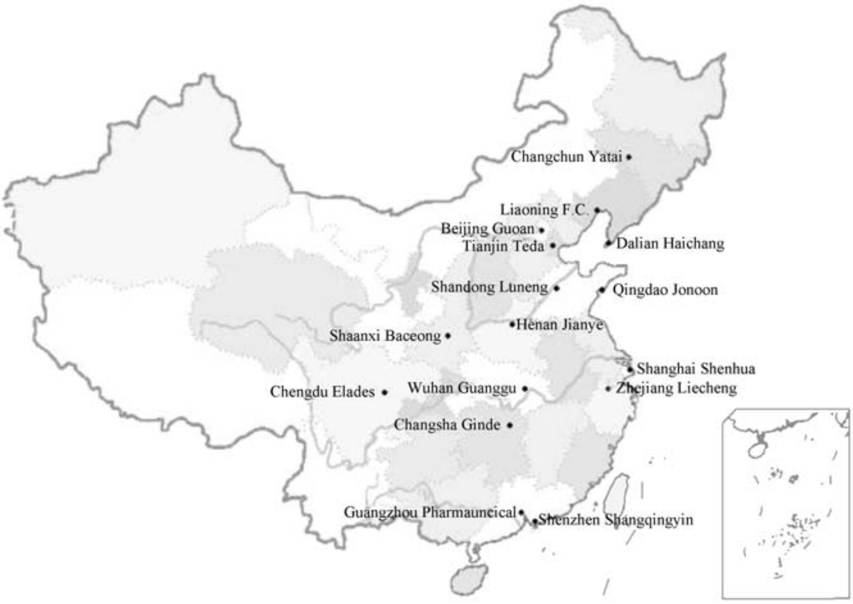

Figure 1 Locations of the 16 Chinese Super League Teams in 2008. The figure is adapted from Wikepedia.

The 2008 season was due to run from 29 March to 6 December, but Round 9 was postponed because of the major tropical storm, Fengshan. Further two other games were postponed because of the great Chengdu earthquake on 12 May. The Olympics also caused some disruption, but all the replays have been completed at this time.

The data to be analyzed are the results for the first 15 rounds of the 2008 season with as an explanatory for forming predictions the list of all the fixtures for the 2008 season. The data on game results may be found at the sites,

www.soccerpunter.com/soccer-statistics/china,
www.soccerway.com/national/china-pr/super-league,
www.rssf.com/tablesc/china08.html,
csl.sports.sohu.com,
www.goalzz.com/main.aspx?c=3604&stage=1,
www.fifa.com/associations/association=chn/nationalleague/standings.html,

while all the fixtures are given at sports.sohu.com/20080321/n255842227.shtml.

The standings after round 15 are given in Table 1. The column of total points obtained at Round 15 is at the rightmost. Shaanxi is leading. In the table home counts are the left three columns, headed W, T, L, and away are the next three. The points for a team involved in a game are determined by awarding 3 points for a win and 1 for a draw/tie. Notice that the total points are not completely in increasing order. This is because the identifiers were assigned using the standings at an earlier round.

Table 1 Super League results after 15 Rounds

Identifier	Team	GP	W	T	L	W	T	L	Points
1	Shaanxi	15	6	1	0	4	3	1	34
2	Shandong	15	7	1	0	2	4	1	32
3	Shanghai	15	7	1	0	2	3	2	31
4	Beijing	15	4	1	1	2	5	2	24
5	Henan	15	4	2	2	1	3	3	20
6	Tianjin	15	3	3	2	3	2	2	23
7	Zhejiang	15	1	3	2	4	3	2	21
8	Guangzhou	15	4	1	3	2	1	4	20
9	Chengdu	15	2	5	1	1	3	3	17
10	Changchun	15	2	4	0	2	2	5	18
11	Quingdao	15	2	4	2	1	2	4	15
12	Changsha	15	2	4	2	1	2	4	15
13	Dalian	15	3	6	0	0	1	5	16
14	Wuhan	15	2	2	4	0	3	4	11
15	Shenzhen	15	1	5	2	0	2	5	10
16	Liaoning	15	2	0	4	0	4	5	10

3 Some previous work on football/soccer statistics

There are now quite a number of papers on the analysis of football data. In particular one can mention [13–15], and the various references in those papers. [14] is a fundamental paper. His data were for the English Premier League and the 1996/1997 season. One of his questions was: "Is Manchester United really the best?" This is another phrasing of Mosteller's question mentioned in the Introduction. Another was, "How can we calculate the probability that a given team will win the Premier League?" Lee worked with the teams' goals scored in a game and assumed the counts to be independent Poisson variates.

By simulation of the model and "determining" the Champion 1000 times, Lee estimated the probability that Manchester United would be the Champion had been 38%. The second team, Liverpool, had the chance of 33%.

Karlis and Ntzoufras[15] worked with bivariate Poisson models for the counts of goals and developed a time series result. Further they considered negative binomial models but did not find them much of an improvement. In their 2008 paper[16] they take the basic variate as the difference of the two teams goals arguing that doing so removes variability in common to the two teams.

Brillinger[5, 6] studied the outcomes of the Norwegian Premier League games of 2003 and Brazilian of 2006 and 2007, respectively, taking the responses to be ordinal-valued (W-T-L) as is the concern here.

4 Ordinal-valued variables

The basic data type in this study is the ordinal-valued, here win, tie, or loss. There are specific numerical values associated with these in football/soccer, 3 points for a win (W), 1 for a tie (T) and 0 for a loss (L). In the work the following model will be employed. It involves a trinomial variate, (corresponding to W, T, or L) and the probabilities $\pi_1(\mathbf{x}), \pi_2(\mathbf{x}), \pi_3(\mathbf{x})$ involving $\mathbf{x}$ an explanatory and parameters to be estimated. With i and j subscripts corresponding to particular teams the model to be employed is

$$\text{Prob}\{i \text{ wins at home playing } j\} = 1 - \exp\{-\exp\{\beta_i + \gamma_j + \Theta_2\}\}, \tag{1}$$

$$\text{Prob}\{i \text{ ties at home against } j\} = \exp\{-\exp\{\beta_i + \gamma_j + \Theta_2\}\} - \exp\{-\exp\{\beta_i + \gamma_j + \Theta_1\}\}, \tag{2}$$

$$\text{Prob}\{i \text{ loses at home against } j\} = \exp\{-\exp\{\beta_i + \gamma_j + \Theta_1\}\}, \tag{3}$$

with the β, and γ effects, both taken to sum to 0, and with $\Theta_2 > \Theta_1$. The effects β, and γ represent home and away effects of the teams. The Θ's relate directly to the overall probabilities of the home team winning at home and losing at home respectively. At the moment the parameters are taken to be constant throughout the season.

It will be shown in the Appendix that the expressions (1)–(3) may be associated with a real-valued latent variable taking values in 3 contiguous disjoint intervals of the real line. This leads to the correspondence with ordinal values. It is also shown in the Appendix that given three trinomial probabilities there may always be associated a latent variate and two cutpoints leading to an ordinal structure.

Following the model (1)–(3) the β's, and γ's have the following implications:
team i tends to win at home if β_i is large,
i tends to lose at home if β_i is small,
i tends to win away if γ_i is small,
i tends to lose away if γ_i is large.

In summary, with the parametrization (1)–(3), relatively speaking team i tends to do well for β_i large and γ_i small. This may be seen directly for the leading team, Shaanxi, in Figure 2.

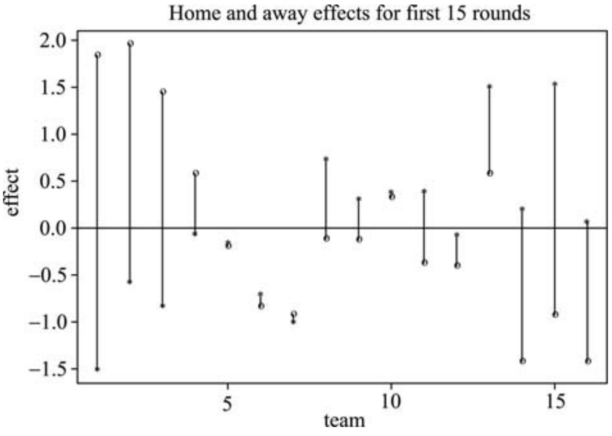

Figure 2 Teams' estimated home and away effects for the first 15 rounds of the season, $\hat{\beta}_i$ and $\hat{\gamma}_i$, "o" and "*" respectively. The numbering is the Identifier of Table 1. Team 1, Shaanxi, is seen to have the second largest $\hat{\beta}_i$ and the smallest $\hat{\gamma}_i$. Team 16, Liaoning, is tied for the smallest β_i and its γ_i is moderately large.

5 Ordinal-valued time series

There are a variety papers on the analysis of time series whose values are categorical. Papers and books that one can mention are [17–22]. Various pertinent computer procedures are described in those papers and could be employed with the football data of concern in this paper. However a loss in efficiency must be anticipated as the present ordinal-valued property does provide information.

The paper [13] stands out as of particular pertinence in some aspects. It studies data from the German Bundesliga for the period 1966–1987. A latent variable cutpoint model is employed. A states space variant is set up and used to track the performance of 6 teams during the 1966–1987 period. It allows time varying "abilities" modelling them by both a random walk and a local linear trend and estimating them by a Kalman filter algorithm. The game results are considered statistically independent given the temporal history.

Brillinger[4] studied National Hockey League outcomes for one team, the Toronto Maple Leafs, for the 1993–4 season. He employed a latent variable cutpoint model. The result for the previous game was included, but not found to be significant.

6 Results

The model (1)–(3) was fit to the W-L-T data for the first 15 rounds of the 2008 Super League Championship. It was assumed that the individual games in a given round were statistically independent of each other. In defence of that assumption it can be remarked that when there is a chance that a team's behavior varies when the results of two games are important those games are played at the same time in major tournaments.

The parameters are estimated by maximizing the likelihood, having in mind that one would like to estimate the final standings for the 2008 season, the probability of each team being in the top four places, and studying the number of points each team might end up with.

A sequence of models was run as follows:

1. Each team has the same home and away effects, i.e. β_i, γ_j assumed 0 in (1)–(3);
2. Home effects assumed;
3. Away effects added next.

The corresponding changes in deviance are given in Table 2. The model improves at each step. The smallest probvalue for improvement is when the away effect is added in. The Θ's, the cutpoints, may be interpreted as providing an overall home advantage. The β's may be related to particular teams' strength playing at home beyond the overall home advantage. The γ's similarly reflect individual teams' strengths when playing away.

Table 2 Changes in deviance

Variate	deviance change	degrees of freedom	probvalue
Cutpoints, Θ	19.487	2	0.000059
Team home strength, β	37.916	15	0.00093
Away team strength, γ	30.429	15	0.01047

Figure 2 displays, team by team, the estimates of β and γ denoted by "o" and "*" respectively. In the fitting the effects of both β and γ, are taken to sum to 0, and this is reflective of how the 0 line passes through the points. In considering Figure 2, following the discussion of Section 4, one wants the home effect, β_i, to be large and the away effect γ_i to be small if one wants team i to be successful. The deviance analysis above provided evidence for the existence of these effects. The leader, team 1, is Shaanxi while the lowest, team 16, is Liaoning. One notes that the left hand lines have "o" at the top while the right hand ones have it at the bottom.

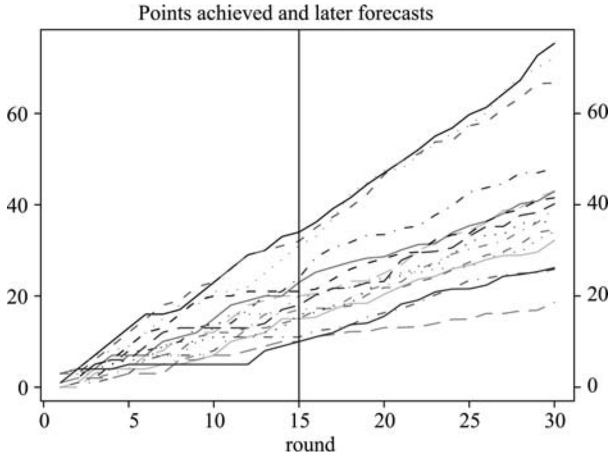

Figure 3 Forecasts of teams' final points in 2008 using the data available up to 6 September, Round 15, and then employing the predicted results thereafter. The vertical line indicates when the available results end and the predicted ones start.

7 Uses and applications

Using the results of fitting the model (1)–(3) it is possible to create probabilistic forecasts of the final points of teams in a league in real time that is, as the season progresses. One fits the model using the data up to Round r. Then one uses the parameter estimates obtained earlier and the list of remaining scheduled games to estimate probabilities of each team's wins, ties, and losses in all its remaining games. At that point one can estimate the expected points at any given future times, particularly the last day of the season. Specifically for a chosen team one adds 3 times the win probability and 1 times the tie probability. The answer provides an estimate of the expected number of points for that team.

Figure 3 shows the results of doing this for each of the Super League teams using the first 15 rounds (20 July 2008) of data, i.e., all the data available at the time the computations commenced.

The vertical line at Round 15, distinguishes the known point totals from the predicted. One sees that the top team, Shaanxi, stays top much of the time, in both the first 15 rounds and then with projected points the last 15 rounds.

One notes, from the results Round 15 and earlier, that one team Liaoning lost 8 games in a row and then were projected to begin winning at an approximately steady rate.

If desired one could work out standard errors for positions along the paths and the final ones.

A simulation was carried out to estimate the theoretical probabilities of the various teams being in the top four places. These teams go on to the Asian Cup, while the champion goes on to the Asian Champions League. Table 3 gives the results based on 25000 simulations. It is interesting to note the results for teams 8 through 13. Because of the vagaries of the scheduling they still have a chance of going onto the Asian Cup. The teams with 0 may also, but it can not be said for sure on the basis of these simulations.

Perhaps Beijing does have a chance of being the champion. Perhaps more simulations will turn up a sequence of results allowing it. A surprising result of [1] shows the problem of determining whether a given team still has a chance of being champion is NP-hard.

Table 3 Simulation results

Identifier	Team	Champion count	Top 4 count
1	Shaanxi	24077	25000
2	Shandong	251	25000
3	Shanghai	672	25000
4	Beijing	0	17075
5	Henan	0	3137
6	Tianjing	0	3385
7	Zhejiang	0	963
8	Guangzhou	0	21
9	Chengdu	0	33
10	Changchun	0	353
11	Qingdao	0	7
12	Changsha	0	17
13	Dalian	0	9
14	Wuhan	0	0
15	Shenzhen	0	0
16	Liaoning	0	0

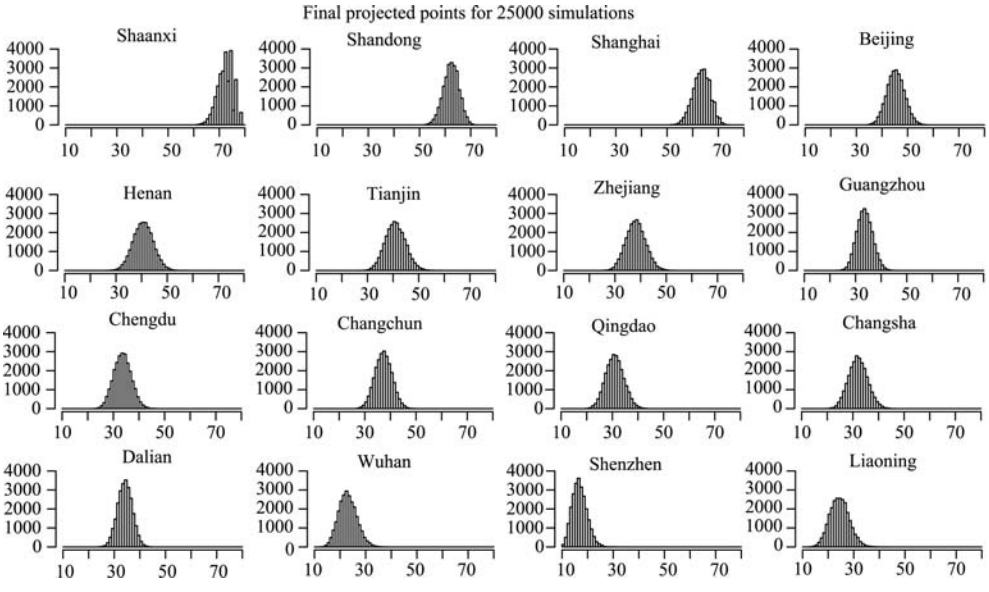

Figure 4 Histograms of the simulated final points for the indicated teams.

In particular the simulation can be used to estimate the probability of the leader Shaanxi being the ultimate, and true, champion. The number of runs in the simulation was 25000. It turned out that Shaanxi was champion in 24077 of the 25000 runs, or in 96.3% of the cases. This is the estimated chance of being champion given the probabilities of win, tie, or loss home and away estimated from the data. It is worth emphasizing that the simulation used the estimates from the first 15 rounds as parameter values. This leads to dependence among the cases. It is to be noted that an implicit assumption for the results to be plausible is that the estimated values are reasonable.

Continuing to make use of the results of the simulation, one needs to consider their variability. Figure 4 provides histograms of the simulated points total for each team. They are ordered as in Table 1 running from left to right across the rows. One sees the histograms sliding from the right to the left with considerable spread. One sees that the Shaanxi histogram appears narrowest and tallest of them all. It is further skewed to the left. Shaanxi does reach 80 points in some cases, but Shandong does not get above 70 much.

One might use the results of the simulations to study the distribution of the number of ties in each round. Further there are opportunities for estimating probabilities related to gambling.

8 Model fit

Next the appropriateness of the model (1)–(3) is considered. Figure 5 graphs the fitted total points versus the actual employing the data for the first 15 rounds. Shaanxi is at the top with 34 points. The points are seen to cluster around the 45 degree line through the origin. The fit appears reasonable.

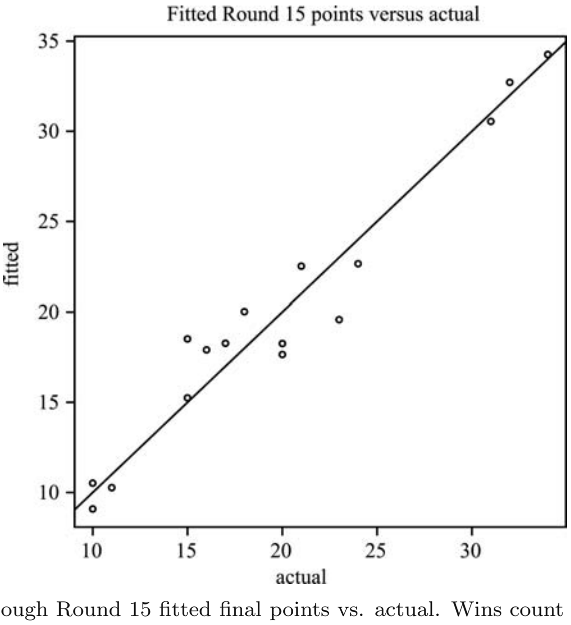

Figure 5 Through Round 15 fitted final points vs. actual. Wins count 3 points, ties 1.

In another study for each of the 96 W-T-L counts of Table 1, the residuals

$$\sqrt{4\text{count} + 1} \;-\; \sqrt{4\text{fit} + 1} \tag{4}$$

were computed and are plotted against their team identifier in Figure 6. Historically the square root transform has been common in work with counts. It is used with the multiplier 4 and the additive 1 for then the variance is approximately 1 in many cases.

The residual values in the figure range approximately between −1.75 and 1.17. No departures from 0 appear large, but there is something of an indication of increasing variability as one moves down the list of teams.

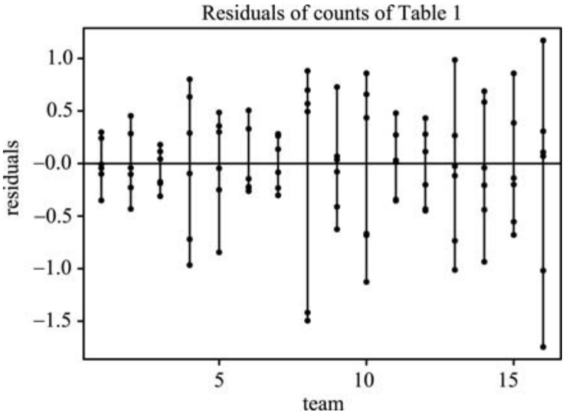

Figure 6 Plots of residuals, as defined at (4).

The appropriateness of expression (1) may be assessed empirically. One splits the predictor values $\hat{\beta}_i + \hat{\gamma}_j + \hat{\Theta}_2$ into disjoint contiguous cells. Then one works out the proportions of times the home team won for the cases in those cells. Next one plots the proportions against the mean point of their cells. The result is Figure 7. Also the approximate marginal 95% confidence limits are displayed via the polygon. Finally the theoretical curve, $1 - \exp\{-\exp\{\eta\}\}$ is added to the plot. The fit is not unreasonable.

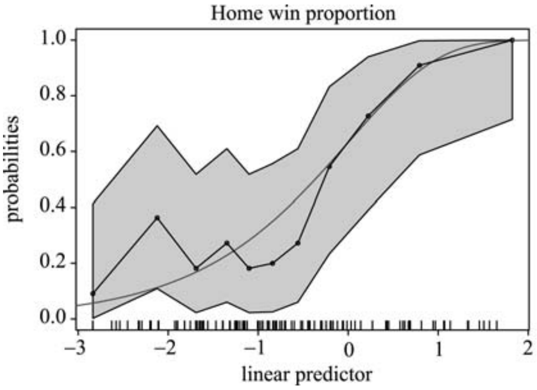

Figure 7 The points are the estimated win at home probabilities. The polygon gives approximate marginal 95% confidence limits. The smooth curve is the assumed form of (1).

9 Including time

In this section two other explanatory variables are considered round number and result of most recent game.

One can use round number as a proxy for date. A variate such as the number of ties in a round could be plotted against the round number as a method to look for trend or change in some distributional characteristic. However, it may be argued in the former case that since a consideration in the scheduling might have led to the teams' positions being assigned randomly, this was not likely to happen. The inclusion of β's and γ's might take out much of any apparent trend effect.

A second exploration studied whether the two teams previous games results (win, tie, or loss) might effect the result. To study this, two 3 level factors were constructed and included in the model. The data may be put in convenient form for such an anlysis. An example of the data for a round in the 2008 season is given in Table 2. The home team's name and result are in columns 1 and 2. The visiting team name is in the 4th column. Columns 3 and 5 give respectively the home team's and visiting team's preceding game results. For example Shanghai was at home in Round 13 playing visiting Dalian. Shanghai won. In Round 12 Shanghai was away playing Changchun and tied. In the same vein in Round 12 Dalian was at home and tied against Zhejiang. One has in mind that a team's most recent result might influence the outcome of the current game.

Table 4 An example of the data of Round 13. Column 2 gives the home team. Column 1 gives their result given the visitor listed in column 4. Column 3 gives their result in the previous round, 12. Column 4 gives the visitor and column 5 their result in Round 12

Result	home	previous	visitor	previous
W	Shanghai	T	Dalian	T
W	Henan	T	Guangzhou	T
T	Changsha	L	Beijing	W
W	Shandong	W	Wuhan	T
L	Tianjing	W	Changchun	T
W	Qingdao	L	Liaoning	L
L	Shenzhen	L	Zhejiang	T
L	Chengdu	T	Shaanxi	W

Including that term the deviance dropped by 3.528 for a change in degrees of freedom of 4. An effect did not appear to be present and none was included in the remaining computations.

10 Discussion and summary

The work has shown that it is feasible to set down a model for Super League game results partially through a season, to fit it, and then given the schedule of the remaining games project the future results. The work has proceeded by modelling the game outcomes rather than the number of goals teams score.

It was considered whether the round number of a game had an effect on the game's result and also whether the result of a team's previous game improved the fit. In the latter case no effect was noted.

It was shown how the model could be used to project the future final points of the teams and thereby the champion as well as the top four teams.

Simulation was employed to address the problems posed and had the advantages of simplicity and flexibility. Exact computations seemed unnecessary for some probabilities. Implementing and running the simulations took little time.

By analysis of deviance it was learned that playing at home was important for all teams, but noticeably more important for some of them.

Acknowledgements Ryan Lovett and Phil Spector helped with computer questions.

References

1 Kern W, Paulusma D. The new FIFA rules are hard: computing aspects of sports competitions. *Discrete Appl Math*, **108**: 317–323 (2001)

2 Albert J, Bennett J, Cochran J J. Anthology of Statistics in Sports. Philadelphia: SIAM, 2005

3 Mosteller F. Lessons from sports statistics. *Amer Statist*, **51**: 305–310 (1997)

4 Brillinger D R. An analysis of an ordinal-valued time series. In: Athens Conference on Applied Probability and Time Series Analysis II. Lecture Notes in Statistics. New York: Springer-Verlag, 1996, 73–87

5 Brillinger D R. Modelling some Norwegian soccer data. In: Advances in Statistical Modelling and Inference. Nair V J, ed. New Jersey: World Scientific, 2006, 3–20

6 Brillinger D R. Modelling outcomes of the Brazilian 2006 Series A Championship as ordinal-valued. *Braz J Probab Stat*, **22**: 89–104 (2008)

7 Tsu chu. In: The Encyclopedia of World Soccer. Washington: New Republic Books, 1979, 689

8 China, People's Republic. In: The Encyclopedia of World Soccer. Washington: New Republic Books, 1979, 130–132

9 Football in China. Wikipedia (2008). //en.wikipedia.org/wiki/Football_in_China

10 Chinese Super League. Wikipedia (2008). //en.wikipedia.org/wiki/Chinese_Super_League

11 Chinese Super League 2008. Wikipedia (2008). //en.wikipedia.org/wiki/Chinese_Super_League_2008

12 Simons R. Bamboo Goalposts. London: MacMillan, 2008

13 Fahrmeir L, Tutz G. Dynamic stochastic models for time-dependent ordered paired comparison systems. *J Amer Statist Assoc*, **89**: 1438–1449 (1994)

14 Lee A J. Modelling scores in the Premier League: is Manchester United really the best? *Chance*, **10**: 15–19 (1997)

15 Karlis D, Ntzoufras I. Analysis of sports data using bivariate Poisson models. *The Statistician*, **52**: 381–393 (2003)

16 Karlis D, Ntzoufras I. Bayesian modelling of football outcomes: using Skellam's distribution for the goal difference. *IMA J Manag Math*, **20**(2): 133–145 (2009)

17 Kaufmann H. Regression models for nonstationary categorical time series: asymptotic estimation theory. *Ann Statist*, **15**: 79–98 (1987)

18 Fahrmeir L, Kaufmann H. Regression model for nonstationary catergorical time series. *J Time Ser Anal*, **8**: 147–160 (1987)

19 Zeger S L, Qaqish B. Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, **44**: 1019–1031 (1988)

20 Harvey A C, Fernandes C. Time series models for count or qualitative observations. *J Business Economic Statistics*, **7**: 407–417 (1989)

21 Davis R A, Dunsmuir W T M, Wang Y. Modelling time series of count data. In: Asymptotics, Nonparametrics and Time Series. New York: M. Dekker, 1999, 63–113

22 Kedem B, Fokianos K. Regression Models for Time Series Analysis. New York: Wiley, 2002

23 McCullagh P, Nelder J A. Generalized Linear Models. 2nd ed. Boca Raton: Chapman and Hall/CRC, 1989, 153–154

Appendix

Note 1.

The modeling and computations follow the discussion in [23]. It presents a common way to create models involving ordinal-valued random variables. Consider a random variable, Y, taking on the ordinal values $k = 0, 1, \ldots, K$. Suppose there exists a random variable Z with cumulative distribution function F, an explanatory variable $\mathbf{x}$, a coefficient β, and cutpoints $-\infty = \delta_0 < \delta_1 < \delta_2 < \cdots < \delta_K = \infty$ such that

$$Y = k \quad \text{if} \quad \delta_{k-1} < Z - \beta^T \mathbf{x} < \delta_k.$$

In the soccer case with $K = 3$, Y represents the ordinal level of the home team's category, Z may be thought of as related to the playing-at-home effect, and the relative strengths of the two teams unmodelled by the explanatories in $\mathbf{x}$ while the δ's may be thought of as an overall bonus given any team playing at home.

In general one can set down,

$$\text{Prob}\{Y \leqslant k\} = \text{Prob}\{Z \leqslant \delta_k + \beta^T \mathbf{x}\} = F(\delta_k + \beta^T \mathbf{x}),$$

say.

If Z has the extreme value distribution,

$$F(y) = 1 - \exp\{-\exp\{y\}\}, \qquad -\infty < y < \infty, \tag{5}$$

then in the present context of 3 categories, W-T-L,

$$\text{Prob}\{Y = W\} = 1 - \exp\{-\exp\{\delta_1 + \beta^T \mathbf{x}\}\}$$

and

$$\text{Prob}\{Y = W \text{ or } T\} \ = \ 1 - \exp\{-\exp\{\delta_2 - \beta' \mathbf{x}\}\}$$

and one is led to expressions (1)–(3) above by choice of β^T and the $\mathbf{x}$.

The assumption of the existence of a latent variable and cutpoints is not restrictive. Suppose one is given trinomial probabilities π_1, π_2, π_3. There is a latent variate U and cutpoints $\eta_1,\ \eta_2$ and an ordinal variate Y, say, taking on the values $1, 2, 3$ with $\text{Prob}\{Y = i\} = \pi_i$ and

$$\begin{aligned} Y &= 1 \quad \text{if} \quad U < \eta_1, \\ Y &= 2 \quad \text{if} \quad \eta_1 < U < \eta_2, \\ Y &= 3 \quad \text{if} \quad \eta_2 < U. \end{aligned}$$

Take U to be uniform of $[0, 1]$ and $\eta_1 = \pi_1,\ \eta_2 = \pi_1 + \pi_2$. If one writes $Z = \log(-\log(1 - U) - \beta^T \mathbf{x})$ one is led to the model (1)–(3) employed in this paper.

Note 2.

The computations proceed by defining a factor with 16 levels whose entries correspond to teams playing at home, and a second with 16 levels whose entries correspond to teams playing away and a third factor, with two levels, corresponding to δ above.

In the maximum likelihood computations the response vector is taken to be the game results for the home team.

The constraint $\delta_2 > \delta_1$ is made explicit in the computations by writing

$$\delta_2 = \log(\mathrm{e}^{\delta_1} + \mathrm{e}^{\psi})$$

and estimating ψ initially, instead of δ_2.

Comment on the paper "An analysis of Chinese Super League partial results", by David R Brillinger

LEE Alan

Department of Statistics, University of Auckland, Private Bag 920719, Auckland 1142, New Zealand
(email: lee@stat.auckland.ac.nz)

This paper gives a nice analysis of an interesting data set, and develops a method that can be applied to a variety of round-robin sports competitions. The paper models the result of a game between two teams as a trivariate ordinal outcome (Win/Draw/Lose) and thus removes the need to model the actual scores explicitly. Modeling scores exactly has several disadvantages compared to the present approach. First, for only a few sports (notably association football) is modeling individual team scores by a simple distribution feasible. The difficulty is particularly acute for sports where the total team score is made up of different components, for example in American Football where there are touchdowns, field goals, safeties and points after touchdown, all worth different numbers of points. Similar difficulties occur with Rugby Union and Rugby League. The distribution of scores in these cases is rather granular (for example, in rugby league it is unusual for a team to score an odd number of points). Even if these difficulties can be surmounted (for example by considering a half-score, see [1]) then a suitable bivariate distribution must be found. These difficulties make the direct modeling of the probabilities of win, lose and draw quite attractive. A further advantage is that the latent variable model used in this paper can be applied to a wide variety of sports, and can be fitted with standard software.

An interesting issue when modeling a round robin competition is whether or not the form of teams fluctuates over time. This is particularly pertinent when competitions are being modeled over several seasons, as changes in personnel, either through retirements or some form of player draft may have a considerable bearing on the ability of teams to win. A few papers have addressed this issue, using some form of state-space modeling. For example, Fahmeir and Tutz[2] used a model similar to that considered in the present paper, but allowed the regression coefficients to be time dependent, fluctuating under a stochastic model. The model was fitted by maximizing the posterior density of the coefficients. In an analysis of NLF scores, Glickman and Stern[3] used a Gaussian state space model for the differences in team scores, ignoring the granularity of scores mentioned above. The mean differences where expressed as the difference of two "ability parameters" in a similar way to the modeling of probabilities using a latent variable model. The posterior distributions of the parameters were estimated by means of the Gibbs sampler.

Both these papers dealt with several seasons. Papers dealing with a single season have tended to assume that the teams' abilities remain constant over the season. This assumption is particularly crucial in the present paper, which makes an end-of-season prediction based on results in the first half of the season. (Of course, all statistical prediction is based on an assumed stochastic model remaining unchanged into the future.)

As of November 20th, the websites cited in Brillinger's article had data up to the end of 28 weeks of competition for the Chinese Super League. Unfortunately, the Shaanxi team did not live up to its good start to the season. They won 10 of their first 15 games, but could only manage four more wins in the next 13, which dropped Shaanxi to 5th place on the points table. The points table after 28 weeks is shown in Table 1.

Table 1 Chinese Super League points table after 28 weeks

Team	Played	L	T	W	L	T	W	Points
Shandong	28	0	2	12	3	6	5	59
Shanghai	28	0	2	12	3	6	5	59
Beijing	28	1	3	9	3	7	5	52
Tianjin	28	2	4	8	3	5	6	51
Shaanxi	28	3	2	9	4	5	5	49
Changchun	28	0	7	6	8	2	5	42
Qingdao	28	4	5	6	5	4	4	39
Zhejiang	28	3	5	5	5	6	4	38
Guangzhou	28	4	4	7	6	5	2	36
Henan	28	4	4	6	7	5	2	33
Chengdu	28	5	6	4	5	5	3	32
Changsha	28	3	8	3	6	5	3	31
Dalian	28	2	7	5	9	4	1	29
Shenzhen	28	2	8	4	11	1	2	27
Liaoning	28	7	4	3	7	5	2	24
Wuhan	28	14	0	0	14	0	0	0

What impact does this decline in form have on the estimated coefficients? In Figure 1, we plot the estimated home and away parameters for Shaanxi, computed using data for the first 15 weeks, first 16 weeks and so on. Note that the coefficients have been scaled so that both the home and away parameters for Zhejiang are set at zero.

From Figure 1, it is clear that there has been considerable change in the coefficients. The effect of the week-by-week results on the coefficients is quite striking. At the left of the plot, the impact of a home tie and loss makes the home parameters decrease, in contrast to an away win, which results in an increase in the away parameter. In the right hand part of the plot, the four successive losses in weeks 24–27 have resulted in decreases (in absolute value) of both the home and away parameters.

Another difficulty arises when a team is either dominant, winning every game, or totally outclassed, losing every game. Under these circumstances, the maximum likelihood estimates of β and γ for the team do not exist, in the sense of being infinite. The iterative algorithms used to fit the models will return large values for the estimates, and hence large or small values (effectively zero or one) for the probability of a good team losing or a bad team winning.

Finally, we make some comments on common software that can be used used to fit the models. SAS PROC GENMOD can fit the latent variable model, offering a choice between the logistic, probit and complimentary log-log links, as is used in the present paper. The algorithms used to fit the model seemed to perform well and converged in each case, albeit with estimates for Wuhan that are artifacts of the stopping rule. This makes it possible to approximately reproduce Brillinger's results, and was in fact used to create Figure 1.

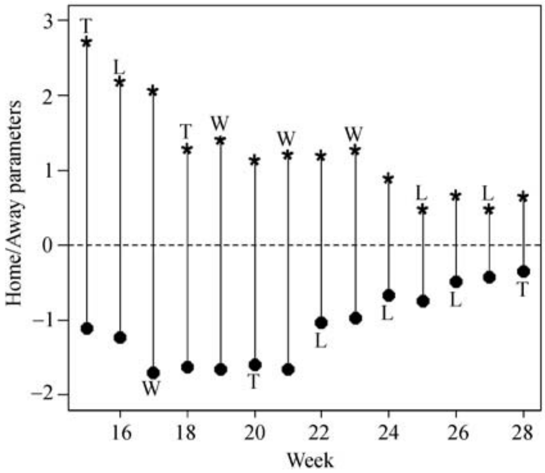

Figure 1 Home and away parameters for Shaanxi, computed for weeks 1 through W, W=15, . . . , 28. The symbol "*" denotes the Home parameter β, and "•" the Away parameter γ.

There is also an R function *polr* in the MASS package, which offers a choice between the probit and logit links. This function is not as robust as the SAS implementation, and it was difficult to get the fitting algorithm to converge in some cases.

In conclusion, the latent variable model offers a simple and versatile method for modeling the outcomes of round-robin tournaments. One must be aware however that the coefficients of the model are quite sensitive to reversals of form and should be interpreted with caution.

References

1 Lee A J. Modelling scores in the Premier League: is Manchester United really the best? *Chance*, **10**: 15–19 (1997)

2 Fahrmeir L, Tutz G. Dynamic stochastic models for time-dependent ordered paired comparison systems. *J Amer Statist Assoc*, **89**: 1438–1449 (1994)

3 Glickman M E, Stern H S. A state-space model for National Football League scores. *J Amer Stat Assoc*, **93**: 25–35 (1998)

Reply to the discussion

BRILLINGER David R

Alan Lee adds to the topic and work of the paper by providing highly insightful discussion and new results. I thank him.

In particular he provides further reasons for working with the derived values, Win-Tie-Loss, rather that actual goal counts themselves.

He refers to the importance of studying temporal behavior adding new results, and he touches on numerical difficulties that can arise.

These topics will be returned to shortly, but first the end of season results. These are given in the table. One notes that the Round 15 projected Champion, Shaanxi, has dropped to sixth position. The eventual Champion was Shandong. One notes that Wuhan withdrew and was given Final points of 0. What the soccer authority did to deal with this was that all Wuhan games after withdrawal were listed as 3 points for the opponent and 0 for them. This meant that there were a number of rounds with essentially 7 games in the "final" data set.

Table 1 Round 15 and Round 30 results

Identifier	Team	Pts	Result	Final Pts
1	Shaanxi	34	.	52
2	Shandong	32	Champion	63
3	Shanghai	31	Champion's League	61
4	Beijing	24	Champion's League	58
5	Henan	20	.	36
6	Tianjin	23	Champion's League	57
7	Zhejiang	21	.	39
8	Guangzhou	20	.	40
9	Chengdu	17	.	32
10	Changchun	18	.	39
11	Qingdao	15	.	39
12	Changsha	15	.	34
13	Dalian	16	.	30
14	Wuhan	11	Withdrew	0
15	Shenzhen	10	.	33
16	Liaoning	10	Relegated	27

Professor Lee remarks on the use of derived, rather than original values. Here is yet another argument for the use of W-T-L. A review of the results from different sites showed that sometimes goal numbers were incorrect, yet the W-T-L result itself was correct. Next Professor Lee comments on temporal aspects. He adds important material and the results of some computations. I have no dispute with this. To continue the temporal discussion the figure below plots the empirical probabilities of wins, ties and losses versus round number. It is meant to assess if there is an overall change in structure as time passes. No overall change is apparent. Professor Lee does demonstrate change for an individual team.

Professor Lee refers to computational difficulties that can arise in certain extreme circumstances, here when a team has 0 wins. In this particular case the 0 is more structural than random. Those games were excluded in the preparation of the figure. However, it seems worth noting that ofttimes the preparers of programs and packages do insert particular values to prevent a program from halting. So really one is led to truncated estimates and these can be useful.

Soccer in China has had its difficulties. This work suggests that statistical regularities do remain for investigation.

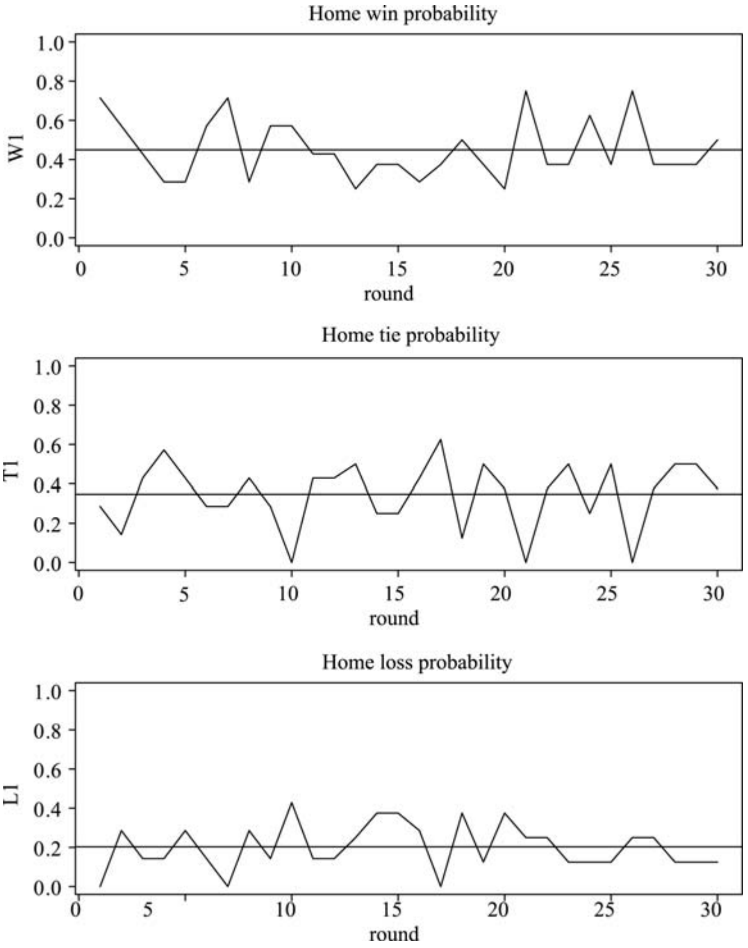

Figure 1 Estimated probabilities of wins, ties, and losses versus round number.