Brazilian Journal of Probability and Statistics
Volume 22, Number 2, Pages 89–104, December 2008
ISSN 0103-0752

MODELLING GAME OUTCOMES OF THE BRAZILIAN 2006 SERIES A CHAMPIONSHIP AS ORDINAL-VALUED

DAVID R. BRILLINGER

"Au football, tout est compliqué par la présence de l'équipe adverse."
Jean-Paul Sartre (1960)

ABSTRACT. Results of the Brazilian Series A football/soccer games are studied for the 2006 season. Other writers have modelled the number of goals a given team scores in a game as a Poisson and then moved on to evaluating the outcome as a win, a tie/draw or a loss via the respective numbers of goals the teams in the game score. This present work takes a different approach. In it the probabilities of win, tie and loss are modelled directly using the outcome results. Model fit is assessed and then there are attempts to improve it by including various explanatories. By simulations using the estimated parameter values it is estimated the probability of São Paulo being the champion, assuming the model, is 63.5% with a standard error of 1.5%. There is some comparative discussion of the Brazilian results with the results for the Norwegian Elite Division in 2003. Then, given the schedule of the remaining 6 rounds, the model is next fit to the results of the first 32 weeks of play in the 2007 Brazilian season. By means of simulation a forecast is then derived of the final number of points each team will have. One finds São Paulo way ahead, but the race to qualify for the Libertadores (top four teams) is quite close. The model may be employed to estimate the chances of many events, for example the distribution of the number of ties on a specified future day knowing which teams are playing.

1. INTRODUCTION

The paper presents an analysis of outcomes, described as win, tie, or loss results for a team involved, of games of the Brazilian Series A 2006 season. Various authors, see the references below, have made a Poisson assumption for the number of goals a team scores in a game. (Sometimes in the data analyses carried out the number of goals is truncated at 5.) A forecast is also made of the final points of the teams in the 2007 employing only the part of the season that has passed on the date this paper was submitted. Round 32 out of the 38 had been completed.

The interest of this paper is in modelling the table, after a certain number of rounds have been completed, taking the basic data as the game outcomes. Here outcome refers

Key words and phrases. Brazilian championship, derived values, final table, football, forecasts, game outcomes, Norway, ordinal data, Series A, simulation.

Received October/2007, accepted December/2007.

to defeat, draw/tie, victory and which team was at home. Thus a team's outcome may be viewed as ordinal-valued.

The model employed, starting with the ordinal assumption for outcome results, is different from the ones starting from the Poisson assumption. One knows of things like weather, referee, cards, injuries could affect the overall count of goals while the "better" team still controls the game and wins. Using the outcomes looks at only whether one teams goal total is greater than the other's and thereby can provide some degree of robustness against extreme numbers of goals. Actually in many of the papers consulted the numbers of goals were truncated at 5. In the Poisson case the numerical difference of goal numbers enters the computation of the parameter estimates. It may be remarked that the collection of outcomes are a sufficient statistic for the final table, which is the item of interest in this work.

Specifically a trinomial model is employed that may be motivated by the assumption of latent variables controlling the final results and having extreme value distributions. The variability of the trinomial is meant to reflect the variability in games due to referee's decisions, player injuries, weather, speaking generally the elements of good luck and bad luck and arbitrariness that affect a game's result.

Various interesting questions arise. Is there a trend as the season progresses? Is there a change? Is there autocorrelation? Which explanatory variables, beyond the teams involved and who is at home, can improve the analysis? Can the approach be employed to develop real time forecasts using the data only up to a current time point? Simulations will be employed to study some of these questions.

The structure of the paper is the following: Introduction, Some previous work on football statistics, Brazilian football, Ordinal-valued variables, Results, Model fit, Explanatory variables, Comparison with Norwegian Elite League results, Uses and applications, Extensions, Discussion and summary, Appendix. In particular Section 9 estimates the probability of each of the teams being in the top 4, i.e. eligible for the Libertadores Cup

2. Some previous work on football statistics

One can mention Fahrmeir & Tutz (1994), Lee (1997), Panaretos (2002), Karlis & Ntzoufras (2003), Fernandez-Cantelli & Meeden (2003), and references in those papers as pertinent previous research. Lee(1997) is a fundamental paper on the subject of analyzing statistically the data of all the teams in a league for a complete season. His data were for the English Premier League and the 1996/1997 season. One of his questions was: "Is Manchester United *really* the best?". 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. The specific model he set down for each team's goals was independent Poissons with means,

$$E\{\textit{home goals for team } i \textit{ playing team } j\} \;=\; exp\{\alpha + \Delta + \beta_i + \gamma_j\} \tag{2.1}$$

$$E\{\textit{away goals for team } i \textit{ playing team } j\} \;=\; exp\{\alpha + \beta_j + \gamma_i\} \tag{2.2}$$

with Δ an overall home effect, with β_i the home effect of team i, with γ_j the away effect of team j, and with the effects standardized by,

$$\sum_i \beta_i, \ \sum_i \gamma_i \ = \ 0 \tag{2.3}$$

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 & Ntzoufras (2003) actually work with bivariate Poisson models for the counts of goals and develop a time series result. Further they consider negative binomial models but don't find them much of an improvement.

3. Brazilian football

The data studied here are from the 2006 and 2007 Brazilian Championships. They were found at the site,

www.soccerway.com/national/brazil/series_a/regular_season/results/1

The 2006 season's final results, listed at the preceding website, are in Table 1 with the column of final points at the rightmost. Home results 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. The champion, São Paulo, ended with 78 points, based on 22 wins and 12 ties. They were well ahead of Internacional who had 69 points.

4. Ordinal-valued variables

The basic data for which the analyses of this study will be carried out are ordinal-valued, win, tie, or loss. The β, γ satisfy (2.3). In Brillinger (1996), National Hockey League outcomes for one team, the Toronto Maple Leafs of 1993-4, and in Brillinger (2006) the outcomes of all the Norwegian Premier League games of 2003 respectively were studied working with the outcomes as ordinal-valued. Of course one can derive the outcomes from the goals scored, but the intention in those papers, as it is here, was to study the effects of taking outcomes as the basic responses.

In the work the following model will be employed. It involves a trinomial variate, (corresponding to win, loss, or tie), with $n = 1$ and the probabilities

$$Prob\{i \ wins \ at \ home \ playing \ j\} \ = \ 1 - exp\{-exp\{\beta_i + \gamma_j + \theta_2\}\} \tag{4.1}$$

$$Prob\{i \ draws \ at \ home \ against \ j\} = exp\{-exp\{\beta_i + \gamma_j + \theta_2\}\} \ - \ exp\{-exp\{\beta_i + \gamma_j + \theta_1\}\} \tag{4.2}$$

$$Prob\{i \ loses \ at \ home \ against \ j\} \ = \ exp\{-exp\{\beta_i + \gamma_j + \theta_1\}\} \tag{4.3}$$

with the β, and γ effects taken to sum to 0 and with $\theta_1 > \theta_2$. The effects β, and γ represent the home and away effects of individual teams. The θs relate to winning at home and losing at home respectively. The β, γ are assumed to satisfy (2.3). The pair (β_i, γ_i) reflect team i's winning advantage. The parameters are assumed constant throughout the season.

TABLE 1. Final results for the 2006 season. São Paulo is the champion. The left W, T, L columns are the home results and the next 3 the away results.

Identifier	Team	W	T	L	W	T	L	Points
1	São Paulo	14	4	1	8	8	3	78
2	Internacional	10	6	3	10	3	6	69
3	Grêmio	13	3	3	7	4	8	67
4	Santos	15	2	2	3	8	8	64
5	Paraná	12	4	3	6	2	11	60
6	Vasco	8	9	2	7	5	7	59
7	Figueirense	9	7	3	6	5	8	57
8	Goiás	9	5	5	6	5	8	55
9	Cruzeiro	10	8	1	4	3	12	53
10	Corinthians	9	3	7	6	5	8	53
11	Flamengo	10	5	4	5	2	12	52
12	Botafogo	10	4	5	3	8	8	51
13	Atlético Paranaense	9	3	7	4	6	9	48
14	Juventude	11	6	2	2	2	15	47
15	Fluminense	7	6	6	4	6	9	45
16	Palmeiras	10	4	5	2	4	13	44
17	Ponte Preta	8	3	8	2	6	11	39
18	Fortaleza	5	5	9	3	9	7	38
19	São Caetano	6	6	7	3	3	13	36
20	Santa Cruz	6	4	9	1	3	15	28

The model (4.1)-(4.3) has 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 (4.1)-(4.3), generally speaking team i tends to do well for β_i large and γ_i small. This will be seen in Figure 1 below.

It is to be noted that for an independent Poisson model the probability of a tie between i at home and j away is

$$\sum_{x=0}^{\infty}(\mu_i\nu_j)^x exp\{-\mu_i-\nu_j\}/(x!)^2$$

with μ_i and ν_j the expected number of goals of the home and away teams respectively. In form this is quite different from expression (4.2) above. It is more complicated analytically.

The model (4.1)-(4.3) may be motivated by the distribution function of the extreme value distribution for the maximum, namely

$$F(y) = 1 - exp\{-exp\{y\}\},\ -\infty < y < \infty \tag{4.4}$$

TABLE 2. Obtained final points and fitted final points for the 2006 season.

Identifier	1	2	3	4	5	6	7	8	9	10
Points	78	69	67	64	60	59	57	55	53	53
Fit	77.16	69.09	66.72	65.74	56.18	63.29	59.62	53.34	53.48	53.45
Identifier	11	12	13	14	15	16	17	18	19	20
Points	52	51	48	47	45	44	39	38	36	28
Fit	51.38	49.33	46.66	49.03	46.65	41.56	38.31	37.30	37.40	27.68

see Note 1 in the Appendix. To fit the model Pregibon (1980)'s trick is employed with the complimentary log- log as the link function, see McCullagh & Nelder (1989), p. 185. Other references include: Läärä and Matthews (1985), and Brillinger et al (2001). Maximum likelihood estimates are derived by employing the function glm() of the statistical package, R, see Ihaka & Gentleman (1996).

The computations proceed by defining a factor with 20 levels which correspond to whom is playing in a game at home, and another with 20 levels which correspond to whom is playing away and a third, with two levels, corresponding to θ above. The constraint $\theta_1 > \theta_2$ is made explicit by writing

$$\theta_1 \;=\; log(e^{\theta_2} + e^{\psi})$$

and estimating ψ initially, not θ_1.

Kedem & Fokianos (2003) have some discussion of the analysis of ordinal-valued time series.

5. RESULTS

A basic goal of this research is to model the final table of standings for the 2006 season, and to study the property of becoming the champion. The model (4.1)-(4.3) is fit to the data of game results assuming the outcomes of the various games are statistically independent.

Figure 1 shows the estimates of β and γ denoted by "o" and "*" respectively. In the fitting the effects are taken to sum to 0. Notice the "o" at the tops of the lines for the better teams and at the bottoms for the lesser. One notes that the champion, team 1 São Paulo, has the largest difference $\hat{\beta}_1 - \hat{\gamma}_1$. One also notes that team 4, Santos, has the largest home effect. The values $\hat{\beta}_i - \hat{\gamma}_i$ fall off more or less monotonically as one moves to the bottom of the table.

The results of the maximum likelihood estimation may be put to various uses, for example to estimate the theoretical numbers of wins-losses-ties, home and away. For São Paulo, see Table 2, the fitted number of points is 77.16 while the actual number is 78. This number being an estimate it is of interest to study its uncertainty, noting particularly that the second place team, Internacional, had 69 points at the end of the season. The fit is investigated in Section 6 below. Table 2 provides the fitted final points and the estimated expected numbers of final points for each team.

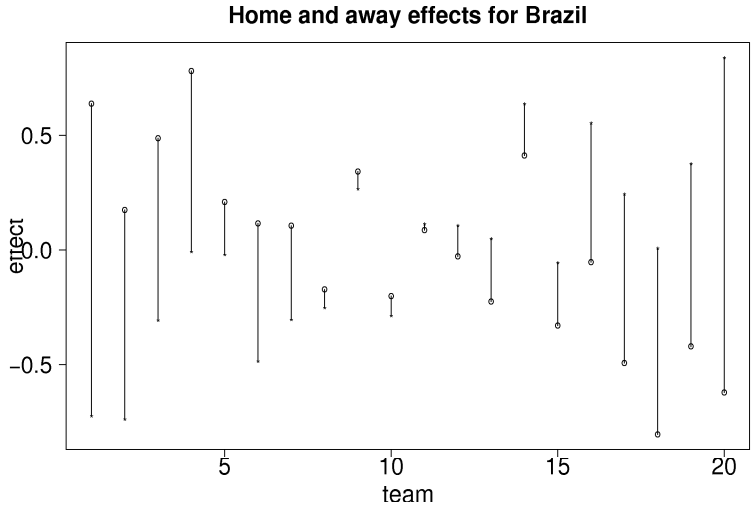

FIGURE 1. Teams' estimated home and away effects, $\hat{\beta}_i$ and $\hat{\gamma}_i$, "o" and "*" respectively for the 2006 season. The numbering is the Identifier of Table 1. Team 1, the champion, is seen to have the largest $\hat{\beta}_i - \hat{\gamma}_i$.

To investigate the uncertainty assume trinomial distributions for the game outcomes, take the outcomes of the various games to be statistically independent, and proceed via the representation,

$$\Sigma_{j\neq 1}\ (3[\delta_{Wj} + \delta_{wj}] + [\delta_{Tj} + \delta_{tj}])$$

for the particular case of São Paulo whose identifier is 1. Here the δ's take the values 0 and 1. The capital letters $W,\ T$ refer to wins and ties in home games and the $w,\ t$ refer to the aways.

To estimate the variance one can use the fact that for a multinomial, with $n\ =\ 1$, the variances are $\pi_j(1-\pi_j)$ while the covariances are $-\pi_j\pi_{j'}$ for $j\ \neq\ j'$. Substituting the estimates $\hat{\pi}$ one obtains 7.004 for the estimated standard deviation of the number of points of São Paulo. This seemed rather large so a simulation was carried out with 1000 runs. It lead to a standard error of 6.774. On reflection the value was large in part because the multiplication by 3 increases the estimates' spread.

That simulation could also be used to estimate the probability of São Paulo being the champion. It turned out that São Paulo was champion in 635 of the 1000 runs, i.e. the estimated probability of being champion was 63.5% with two standard errors of 3.0%. (When two teams tied for the largest number of points, the champion was selected between them randomly. See Note 2 in the Appendix.) This is the estimated chance of being champion given the probabilities of win, tie, or loss home and away estimated from the data.

Figure 2 provides boxplots of the 1000 simulated total points that each team received. It shows substantial variability for both the top and bottom teams. The champion team 1,

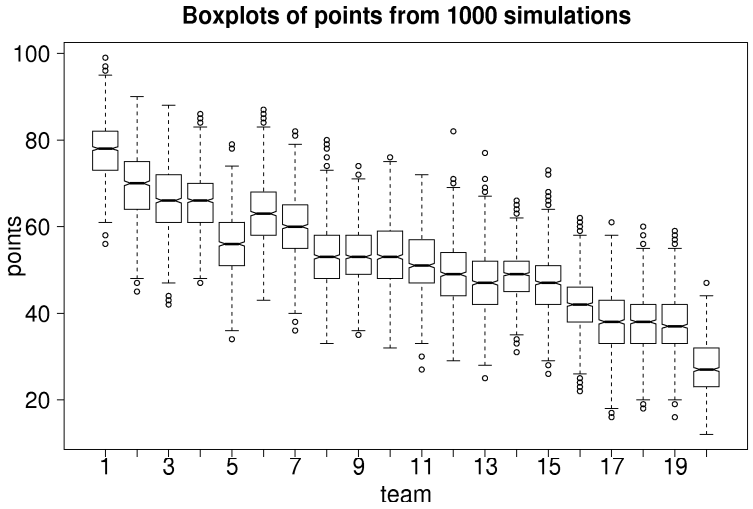

FIGURE 2. Notched boxplots of teams' simulated final points derived from 1000 simulations. For the team name corresponding to the identifier on the x-axis see Table 1.

São Paulo, rests far above the others quite generally. As Figure 1 shows its home effect, β, is high and its away effect, γ, is low. As indicated earlier, both of these lead to an increased chance of winning in formula (4.1).

The simulation was also carried out awarding 2 points for a victory. A change to awarding 3 points had been made in 1981. The thought was that it would lead an increase in attacking football, to more goals, and to reducing the numbers of ties. The results of Figure 2 were not much changed.

6. MODEL FIT

Next the appropriateness of the model (4.1)-(4.3) was checked. Figure 2 suggested that team 5, Paraná, might be an outlier, as did Table 2, but further investigations did not support that idea. Figure 3 graphs the fitted points versus the actual having assumed 3 points for a win. The points cluster around the 45 degree line through the origin. The fit appears reasonable.

For each of the 120 counts of Table 1, the residuals

$$\sqrt{4 * count + 1} \; - \; \sqrt{4 * fit + 1} \tag{6.1}$$

were computed and plotted against their team identifier in Figure 4. The square root transform is common in work with counts and its variance is approximated by 1. The residual values in the figure range approximately between -1.5 and 1.5 . No departures from the model stand out.

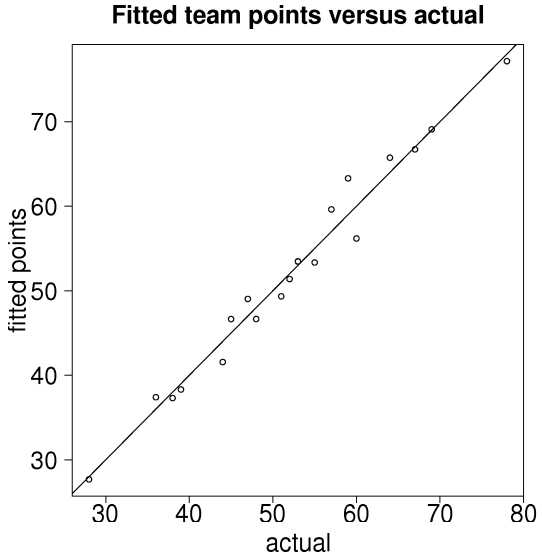

FIGURE 3. For the 2006 season fitted final points vs actual. Wins count 3 points.

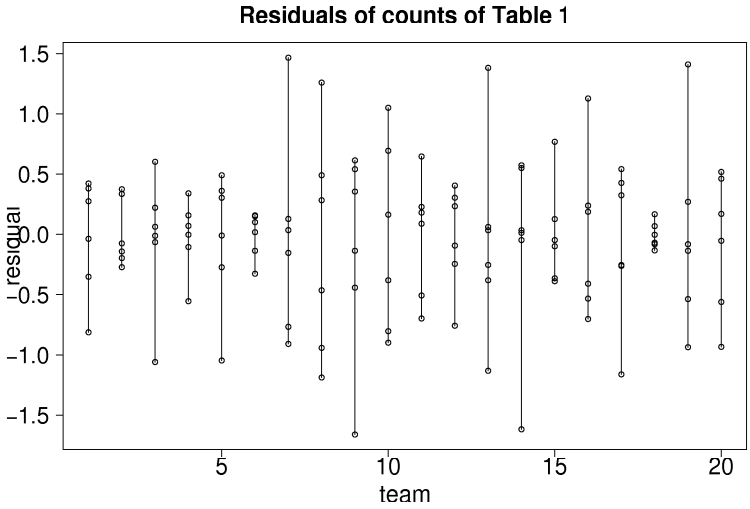

FIGURE 4. Plots of differences, as defined at (6.1), between observed counts given in Table 1 and fitted counts.

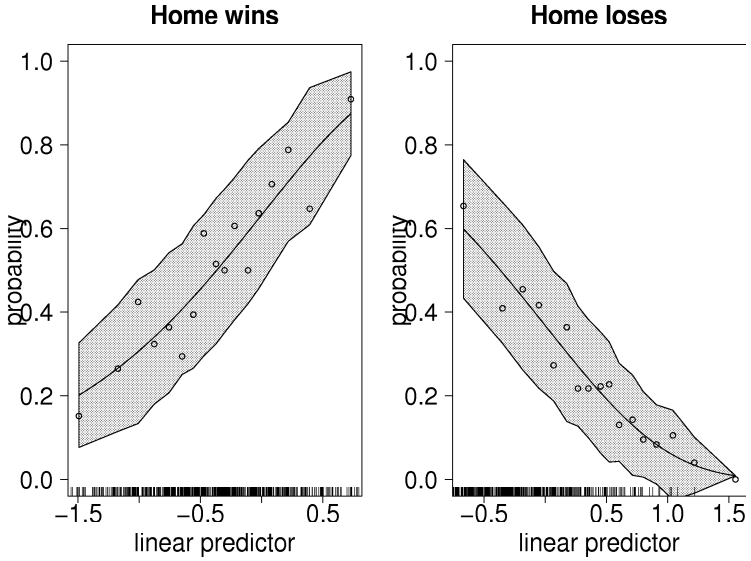

FIGURE 5. For the 2006 season the lefthand plot concerns the model for a team to win at home, and the righthand for a team losing at home, see the text.

The extreme value distribution of the maximum, corresponding to the inverse complimentary log-log link, appears in expression (4.1) and was employed in the fitting. Figure 5 provides a check on assumptions (4.1) and (4.3). It graphs the empirical probability, the proportion of 1's in cells based on the estimated linear predictor, $\hat{\beta}_i + \hat{\gamma}_i + \hat{\theta}_2$, against linear predictor cells. The lefthand panel of Figure 5 is the curve is $1 - exp\{-exp\{\eta\}\}$. The linear predictor, η, for the right hand panel is $\hat{\beta}_i + \hat{\gamma}_i + \hat{\theta}_2$ while, following (4.3) the curve is $exp\{-exp\{\eta\}\}$. The shaded areas are meant to provide approximate marginal $\pm$ 2 standard error limits. These are computed as $2\sqrt{\hat{p}(1-\hat{p})/n}$ where $\hat{p}$ is an empirical proportion and n the number of cases on which it was based. The plotted points and curves seem reasonable, but there is an indication that the 2 standard error limits are too broad. The rug plots along the bottom of the figures indicates the locations of the linear predictor values.

7. EXPLANATORY VARIABLES

In this section two other explanatory variables are considered beyond the previous home and away ones.

A particular type of residual plot may be used to look for dependence on the game date. Normal residuals are employed and the results are given in Figure 6. Normal residuals are useful for binary data and are designed to have approximate standard normal distributions if the model is reasonable. Specifically if Y is a 0-1 valued random variable

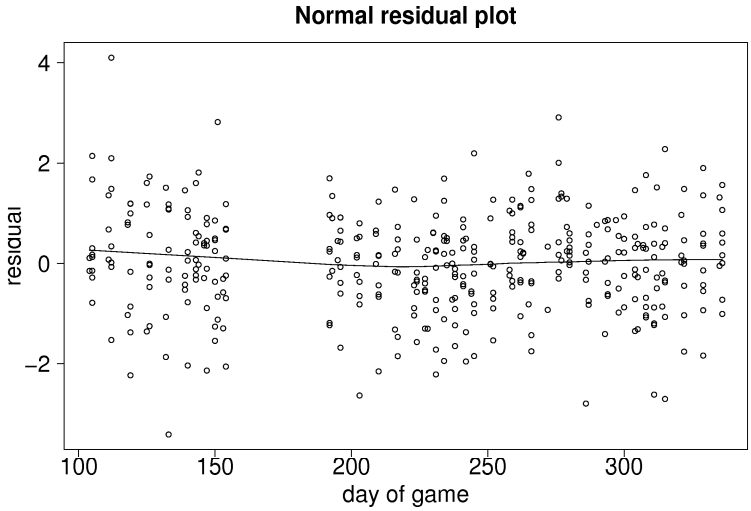

FIGURE 6. Normal residuals vs julian day of 2006 for the case of home wins. The gap corresponds to the World Cup break.

with $Prob\{Y = 1\} = \pi$, and $\hat{\pi}$ is an estimate of π to be examined the they are computed as

$$\Phi^{-1}(\hat{U}_1(1 - Y) \ + \ \hat{U}_2 Y)$$

where $\hat{U}_1$ and $\hat{U}_2$ are independent uniform on $(0, 1 - \hat{\pi})$ and $(1 - \hat{\pi}, 1)$ respectively, see Brillinger (2006). Here the residuals are plotted against the day of the game to look for a time trend taking place during the season. Figure 6 does show a bit of bending down from the start, but it is not extreme. The gap in the plot between days 154 and 192 corresponds to the period when the League ceased play as the World Cup was taking place.

A second explanatory studied was a team's result (win, tie, or loss) in the previous game. When this was done the deviance dropped by 6.4 with the degrees of freedom increasing by 4. Mostly the away results were affected by the previous loss results, but the effect was not dramatic.

8. Comparison with the Norwegian Elite League results

Some previous data analyses, of the nature of those of this paper, were carried out using data for the 2003 season of the Norwegian Premier Division, Brillinger (2006). This present paper carries out some further analyses, for Brazilian data.

The Norwegian league has 14 teams that play each other twice, once at home and once away. The analyses in this paper may be used for comparing the two leagues. Table 3 lists the Norwegian teams in the order of the final standings in 2003. One notes that Rosenborg ended as the champion, far ahead. One also notes that Bryne, at the bottom of the table, lost all their away games.

TABLE 3. The end of season's results for the Norwegian Premier League in 2003. The W-T-L columns on the left refer to home games and the next three to away games.

Team	Identifier	W	T	L	W	T	L	Points
Rosenborg	1	9	2	2	10	2	1	61
Bodo-Glimt	2	7	2	4	7	3	3	47
Stabaek	3	6	4	3	5	5	3	42
Odd-Grenland	4	6	4	3	5	1	7	38
Viking	5	6	3	4	3	7	3	37
Brann	6	7	1	5	3	6	4	37
Lillestrom	7	7	4	2	3	3	7	36
Sogndal	8	7	4	2	2	4	7	35
Molde	9	6	2	5	3	2	8	31
Lyn	10	4	3	6	4	3	6	30
Tromso	11	4	4	5	4	1	8	29
Valerenga	12	4	5	4	2	5	6	28
Aalesund	13	4	5	4	3	2	8	28
Bryne	14	7	1	5	0	0	13	22

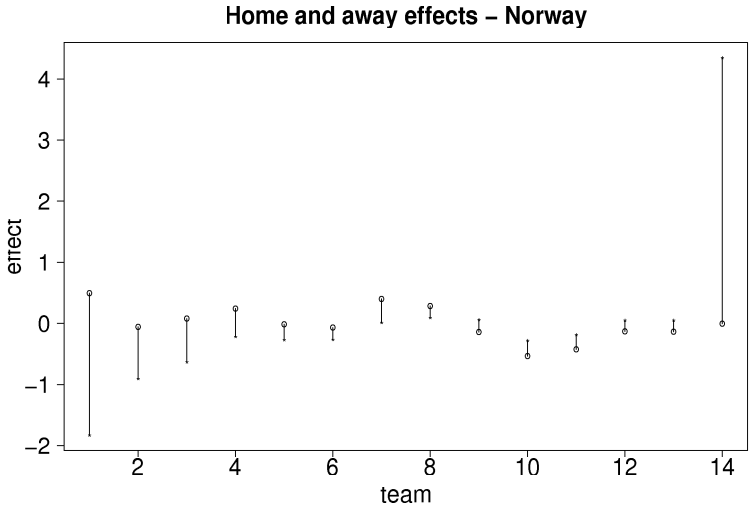

FIGURE 7. Norwegian teams' home $\hat{\beta}$ and away $\hat{\gamma}$ effects denoted by "o" and "*" respectively. For the teams names see Table 3. The team's are ordered by standing at the end of the season with the better from the left.

Figure 7, displays the home and away effects, and may be compared with Figure 1 above for Brazil. The winner has high $\hat{\beta}$ and low $\hat{\gamma}$ as in the previous discussion. The high away-value at the left reflects Bryne, losing all its away games during the season. The Brazil values range from -.75 to .75 while, except for Bryne, the Norwegian values range from -1.9 to .75 . Note there are the conditions (2.3) for the effect estimates.

It is interesting that in both Figure 1 and Figure 7 the values $\hat{\beta}_i - \hat{\gamma}_i$ decrease approximately monotonically as one moves from the top of the table to the bottom. From Tables 1 and 3 the % of the Brazilian teams' wins at home is 67.5, while for the Norwegians' it is 60.8 .

9. Uses and applications

Using the results of fitting the model (4.1)-(4.3) 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 time t. Then one uses the parameter estimates obtained and the remaining scheduled games to estimate probabilities of each team's wins, ties, and losses. Then one can estimate the expected points at any given future times, particularly the last day of the season.

Figure 8 shows the results of doing this for each of the Series A teams using the first 32 rounds, (21 October 2007), of data, i.e. all the data available at this time of writing. The results for São Paulo are shown by the thick black line rising and eventually staying above the rest.

If desired one could work out standard errors for positions along the paths as was done for São Paulo's final total earlier. A simulation was carried to estimate the theoretical chance of Sao Paulo's being the champion. They were 24997 times out of 25000 runs. Santos was 2 times, and Cruzeiro 1 time. So some other teams still had a chance of being champion albeit small. It is interesting to note that the problem of determining, during the season, whether a given team can still become champion has been shown to be NP-hard, Kern & Paulusma (2001).

Because Sao Paulo was already so far ahead, when the paper was submitted, it was realized that a more topical question was what are the probabilities of the various teams being in the top four places. These teams go on to the Libertadores Cup. Table 4 gives the results based on 25000 simulations. One notes that Palmeiras and Cruzeiro are passed over by Flamengo and Fluminense as the season ended. One would have lost betting on the maximum likelihood estimates. More explanatories are needed in the model.

There are other questions of interest that can be addressed directly via simulation. For example given the data up to a particular date one might ask for an estimate of the distribution of the number of ties on the last day of the season. As well there are many opportunities for estimating other appropriate probabilities related to gambling.

10. Extensions

The idea of employing random effects to compensate for missing explanatories was investigated in Brillinger (2006) for the Norwegian data. It did not change the results there.

TABLE 4. Round 32 total points for the 2007 Championship are in the third column. Column 4 gives the count of the number of times the team was in the top 4 places in 25000 simulations. Column 5 gives the total points the team had at the end of the season. Note that Flamengo and Fluminense had entered the top 4.

Identifier	Team	Pts	Top 4 Count	Final Pts
1	São Paulo	67	25000	77
2	Palmeiras	54	19613	58
3	Cruzeiro	53	16510	60
4	Santos	52	12089	62
5	Grêmio	51	14057	58
6	Flamengo	49	8572	61
7	Fluminense	48	3524	61
8	Figueirense	45	70	53
9	Botafogo	45	359	55
10	Atlético-PR	45	297	54
11	Internacional	44	19	54
12	Vasco	43	3	54
13	Atlético-MR	43	76	55
14	Náutico	43	10	49
15	Sport	43	1	49
16	Goiás	41	0	45
17	Corinthians	38	0	44
18	Paraná	34	0	41
19	Juventude	31	0	41
20	América-RN	16	0	17

The change from awarding two points for a win to awarding three occurred in 1981 has already been mentioned. The claims made for this included that it would lead to more attacking football and fewer ties. Fernandez-Cantelli & Meeden (2003) investigated the impact of the change and did not find much. They offer some intriguing suggestions for changing the points awarded, including zero points for a tie. It is to be mentioned that Figure 2 did not change much when the 3 points award was changed to 2.

A future study might involve applying the Lee (1997) Poisson approach to the Series data and seeing how the results compare with those of this paper. As already mentioned it involves a different statistical distribution for the outcomes.

11. Discussion and summary

Ordinal-valued quantities provide an interesting data type that arises in widely varying fields. The model (4.1)-(4.3) has been set down to describe the ordinal-valued outcome of a team playing a football game. The model has been assessed in several ways using the Series A data and found reasonable, so far. It was considered whether the date of the

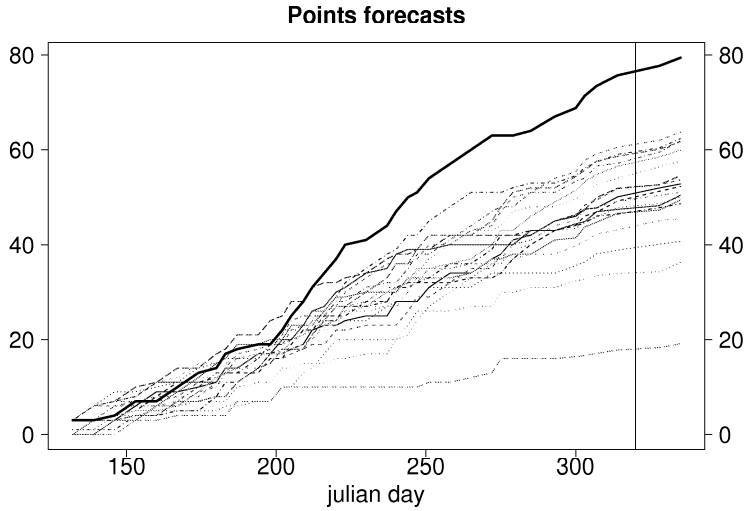

FIGURE 8. Forecast of teams' final points in 2007 using the data presently available, i.e. the first 32 weeks of the season. The heavy dark line is São Paulo. The vertical line indicates when the forecasted results began.

game had an effect on a team's results and also whether the result of a team's previous game improved the fit. Neither changed the results substantially.

It was shown how the model could be fit to an existing stretch of results and then used to project the future final points of the teams and thereby the Champion as well as the top four teams. It was also found, when computing standard errors, that there was substantial variability present in the performance of the teams.

The Lee (1997) question may be reformulated as,

Was São Paulo really the best in 2006?

In other words, were they lucky or not? By simulation the theoretical chance of São Paulo's becoming the Champion was estimated as 63.5%, with two standard error value of 3.0%. This is far greater than Manchester United's 38% as reported in Lee (1997). For the Campeanato the second place team was Internacional. They ended far behind São Paulo and had an estimated winning chance of 14.8%. It is hard to deny that São Paulo was the best in 2006.

Employing simulation had the advantages of simplicity and flexibility. Exact expressions could have been set down for the probabilities, but could have been both complex and messy. Implementing the simulations took hardly any time at all and they could have been left running a long time with no extra intellectual effort.

12. Appendix

Note 1.

The model (4.1)-(4.3) may be motivated as follows. Consider a random variable, Y, taking on the ordinal values 0, 1, 2. Suppose there exists a random variable Λ, an explanatory variable $\mathbf{x}$, a coefficient α, and cutpoints $\theta_1 > \theta_2$ such that

$$Y = 0 \; if \; \theta_1 < \Lambda - \alpha'\mathbf{x} \; leads \; to \; defeat \; for \; the \; home \; team$$

$$Y = 1 \; if \; \theta_2 < \Lambda - \alpha'\mathbf{x} < \theta_1 \; leads \; to \; a \; draw/tie$$

$$Y = 2 \; if \; \Lambda - \alpha'\mathbf{x} < \theta_2 \; victory \; for \; the \; home \; team$$

The variate Λ may be thought of as related to the relative strengths and weaknesses of the two teams If Λ has the extreme value distribution, (4.4), then

$$Prob\{Y = 2\} = 1 - exp\{-exp\{\alpha'\mathbf{x} + \theta_2\}\}$$

and one has expression (4.1) above by choice of α and $\mathbf{x}$. The other two expressions follow similarly.

Note 2.

There is a detail.

Teams are ranked by total points. At the end of the season, the club with the most points is champion. However if the points are equal the number of victories, then the goal difference then the goals scored, then a playoff and lastly a draw as necessary are used to determine the champion. Wikipedia (2007)

In the simulations carried out a tie was resolved by the number of victories and then, if necessary, by a random choice of champion.

Acknowledgements

Lúcia Barroso encouraged the study of São Paulo's 2006 championship season and gave the paper a careful reading. Felipe de Barros helped the author in the preparation of the talk for IME USP. After the presentation Julio Singer mentioned that Santos was known to have a strong home effect. Indeed it may be found in Figure 1. Santos has identifier 4 there and $\hat{\beta}_4$ is the largest home effect in the plot.

The Referee's comments were helpful and thought provoking. They led to restructuring and some other changes.

The research was supported in part by NSF grants DMS-0504162 and DMS-0707157.

References

Brillinger, D. R. (1996), 'An analysis of an ordinal-valued time series, 73- 87', *A*thens Conference on Applied Probability and Time Series Analysis II Lecture Notes in Statistics. Springer-Verlag.

Brillinger, D. R. (2006), 'Modelling some Norwegian soccer data', *A*dvances in Statistical Modelling and Inference, (Ed. V. J. Nair.) World Scientific. 3-20.

Brillinger, D. R., Chiann, C., Irizarry, R.A. & Morettin, P. A. (2001), 'Automatic methods for generating seismic intensity maps', *P*robability, Statistics and Seismology: Journal of Applied Probability **38A**, 189-202.

Fahrmeir, L. & Tutz, G. (1994), 'Dynamic stochastic models for time-dependent ordered paired comparison systems', *J*ournal of the American Statistical Association **89**, 1438-1449.

Fernandez-Cantelli, E. & Meeden, G. (2003), 'An improved system for soccer', *C*hance **27**, 50-53.

Ihaka,R. & Gentleman,R. (1996), 'R: A language for data analysis and graphics', *Journal of Computational and Graphical Statistics* **5**, 299-314

Karlis, D & Ntzoufras, J. (2003), 'Analysis of sports data using bivariate Poisson models', *T*he Statistician **52**, 381-393.

Kedem, B. & Fokianos, K. (2002). *R*egression Models for Time Series Analysis. Wiley, New York.

Kern, W. & Paulusma, D. (2001), 'The new FIFA rules are hard: computing aspects of sports competitions', *D*iscrete Applied Mathematics **108**, 317-323

Läärä, E. & Matthews, J. N. S. (1985), 'The equivalence of two models for ordinal data', *B*iometrika **72**, 206-207.

Lee, A. J. (1997), 'Modelling scores in the Premier League: is Manchester United *really* the best?', *C*hance **10**, 15-19.

McCullagh, P. & Nelder, J. (1989), *G*eneralized Linear Models. Chapman and Hall, London.

Panaretos, V. (2002), 'A statistical analysis of the European soccer Champions League', *P*roc. Joint Statistics Meeting.

Pregibon, D. (1980), 'Discussion of paper by P. McCullagh', *J*ournal of the Royal Statistical Society B **42**, 139.

Sartre, J-P. (1960), *C*ritique de la Raison Dialectique. Gallimard, Paris.

Wikipedia (2007), //pt.wikipedia.org/wiki/Campeonato_Brasileiro_de_2006, 20 October 2007.

Department of Statistics, University of California,, Cidade Universitária, Berkeley, CA, 94720, USA
E-mail address, David R. Brillinger: brill@stat.berkeley.edu