Chapter 1

MODELLING SOME NORWEGIAN SOCCER DATA

David R. Brillinger
Statistics Department
University of California, Berkeley, CA, U.S.A.

E-mail: brill@stat.berkeley.edu

Results of Norwegian Elite Division soccer games are studied for the year 2003. Previous writers have modelled the number of goals a given team scores in a game and then moved on to evaluating the probabilities of a win, a tie and a loss. However in this work the probabilities of win, tie and loss are modelled directly. There are attempts to improve the fit by including various explanatories.

Key words: Binary data; Empirical process; Football; Generalized linear model; Ordinal data; Residuals; Soccer.

1. Introduction

Kjell Doksum has been a steady contributor to the theory and practice of nonparametric statistics and soccer. In former case he has studied the quantile function, probability plotting and, what is most pertinent to this article, the introduction of randomness to ease analyses. In the latter case he has potted lots of goals during his lifetime.

Previous studies have modelled the number of goals a team of interest scores in a soccer game as a function of explanatories such as site of game, opponent, and FIFA rating. References include Lee (1997), Dyte and Clarke (2000), Karlis and Ntzoufras (2000, 2003a, 2003b) and references therein. In our work the respective probabilities of win (W), tie (T), and loss (L) are modeled directly and are examined as a functions of possible explanatories. A reason for employing W, T, L is the thought that the ultimate purpose of a game is to decide a winner. It is felt that the response W, T, L better represents this event than the number of goals scored. The latter may be

inflated by a team's "giving up" or be deflated by a team's moving to a defensive strategy.

To an extent the approach is that taken in Brillinger (1996) for hockey data. The sections of the paper after the Introduction are: Some previous soccer modeling, Norwegian soccer, Ordinal data, Results, Assessing fit, Another model, Uses, Extensions, Discussion and summary.

2. Some previous soccer modelling

There has been previous work on modelling the number of goals scored by each team in a game. For example Lee (1997) employs independent Poissons for the number of home and away goals, with

$$E\{home\ goals\ by\ team\ i\} \ = \ exp\{\alpha + \Delta + \beta_i + \gamma_j\}$$

$$E\{away\ goals\ by\ team\ i\} \ = \ exp\{\alpha + \gamma_i + \beta_j\}$$

respectively, where Δ represents the home effect, β_i refers to team i playing at home and γ_i refers to team i playing away, and j refers to any arbitrary team. On the other hand Dyte and Clarke (2000) employ the expected value

$$exp\{\alpha + \beta U_i + \gamma V_j + \Delta\}$$

where U_i is i's FIFA rating, V_j is j's and Δ is again a home team effect. Karlis and Ntzoufras (2000, 2000a, 2000b) employ the Poisson and bivariate Poisson in their work. In each case the model is used to determine resultant win, tie, loss from the goals scored. In this paper the focus is on the win-tie-loss result as the basic response.

Panaretos(2002) adopts a "game viewpoint" employing explanatories such as: fouls committed, off-sides, and shots on goal. Brillinger (2005) employs mutual information as a measure of the strength of association of the effect of playing at home and the number of goals scored for various premier leagues around the world.

3. Norwegian soccer

Table 1 lists the Norwegian Elite Division teams for the 2003 season, and Figure 3. is a map displaying their locations. One notes the two northerly teams and wonders whether travel and weather might not play important roles in their games. Also listed are identifiers for the map showing locations. (The reason for switching the last five teams from numbers to letters

Table 1 The 2003 Elitserien teams. The identifier provides their location on the map of Figure 3.. The teams are in the order of their 2003 finish

Team identifier	*Team*
1	Rosenborg
2	Bodo-Glimt
3	Stabaek
4	Odd-Grenland
5	Viking
6	Brann
7	Lillestrom
8	Sogndal
9	Molde
10(a)	Lyn
11(b)	Tromso
12(c)	Valerenga
13(d)	Aalesund
14(e)	Bryne

is to have less overprinting in the figure.) The teams are listed in the table in order of their final standings for 2003.

The Elitserien has 14 teams, each playing all the others, home and away. There were 182 games in the 2003 season and the season goes on 26 weeks with a break in the summer. There are 7 games each week and the design is balanced. The data employed in the analyses came from the url http://www.soccerway.com/national/norway/tippeligaen/2003/round-1/results

Table 2 The results of the first week's games as an illustration

Home	Visitor	Result
Rosenborg	Valerenga	1 - 0
Lillestrom	Bodo/Glimt	1 - 0
Aalesund	Tromso	2 - 3
Viking	Bryne	3 - 0
Sogndal	Stabek	2 - 1
Odd_Grenland	Molde	1 - 0
Lyn	Brann	0 - 0

To show the character of the original data the first week's results are displayed in Table 2. Table 3 provides the final 2003 season results. The left hand columns give the at-home results and the right hand the away for the complete season.

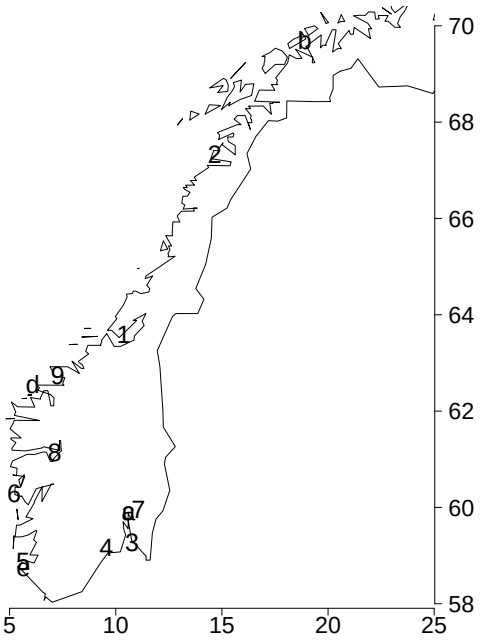

Figure 1 Locations of the 2003 Elitserien Teams. Table 1 lists the team names corresponding to the identifiers. The x-axis is longitude east and the y-axis latitude.

4. Ordinal data

This section motivates and lays out the analysis approach taken in the paper.

4.1. *The cut-point approach*

The random variables of principal concern are ordinal-valued namely *loss, tie, win.* These will be denoted by

$$0,\ 1,\ 2$$

respectively.

A number of different models have been proposed for the analysis of ordinal data. These include: continuation ratio (see Fienberg (1980)), stereotype (see Andersen (1984)) and grouped continuous (see McCu1lagh and

Table 3 The season's results for 2003. The left hand columns are home games and the right hand columns away games

Identifier	W	T	L	W	T	L
1	9	2	2	10	2	1
2	7	2	4	7	3	3
3	6	4	3	5	5	3
4	6	4	3	5	1	7
5	6	3	4	3	7	3
6	7	1	5	3	6	4
7	7	4	2	3	3	7
8	7	4	2	2	4	7
9	6	2	5	3	2	8
10(a)	4	3	6	4	3	6
11(b)	4	4	5	4	1	8
12(c)	4	5	4	2	5	6
13(d)	4	5	4	3	2	8
14(e)	7	1	5	0	0	13

Nelder (1989)). This last is the one employed in the analyses presented.

The approach to be followed starts by supposing that there exists a latent variable, Λ, whose value in some sense represents the difference in strengths of two teams in a game. It further assumes the existence of cutpoints θ_1 and θ_2 such that

$$Y = 0 \text{ if } \Lambda < \theta_1, \quad Y = 1 \text{ if } \theta_1 < \Lambda < \theta_2 \text{ and } Y = 2 \text{ if } \theta_2 < \Lambda$$

so for example

$$Prob\{Y = 1\} = F_\Lambda(\theta_2) - F_\Lambda(\theta_1) \tag{1}$$

where F_Λ is the c.d.f. of Λ. In practice the choice of F_Λ is sensibly based on the subject matter of the problem. The *complimentary loglog* link corresponds to situations in which of an internal variate crosses a threshold. It may be based on an extreme value distribution. In the present context this may be reasonable, with a win for a particular team resulting from the team members putting out maximum efforts to exceed those of the opponent. What is basic though is that its choice makes standard generalized linear model programs available via the Pregibon trick.

The extreme value distribution of the first type is given by

$$Prob\{\Lambda \leq \eta\} = 1 - exp\{-e^\eta\}, \quad -\infty < \eta < \infty \tag{2}$$

One can write

$$log(-log(1 - Prob\{\Lambda \leq \lambda\})) = \lambda$$

and see the appearance of the *cloglog* link. Pregibon (1980) noted that one could employ standard statistical packages in analyses of such multinomial

data when one proceeded via conditional probabilities. Here the distributions involved in the modelling are $Prob\{Y = 2\}$ and $Prob\{Y = 1|Y \neq 2\}$.

Explanatory variables, x, may be introduced directly by writing

$$\Lambda \ = \ \mathrm{E} \ + \ \beta' x$$

where E has the standard extreme value distribution. Now (1) becomes

$$F_{\mathrm{E}}(\theta_2 - \beta' x) \ - \ F_{\mathrm{E}}(\theta_1 - \beta' x).$$

4.2. *Some formulas*

To begin consider $Prob\{Y = 2\}$, as opposed to $Prob\{Y \neq 2\}$, and $Prob\{Y = 1|Y \neq 2\}$, as opposed to $Prob\{Y = 0|Y \neq 2\}$. The response is binary in each case. In the work the following parametrization will be employed,

$$Prob\{Y = 2\} \ = \ 1 \ - \ exp\{-e^{\eta - \theta_2}\}$$

$$Prob\{Y = 1|Y \neq 2\} \ = \ 1 \ - \ exp\{-e^{\eta - \psi}\} \tag{3}$$

with $\eta = \beta' x$. The other probabilities of interest may be obtained from these. The advantage of this parametrization is that θ_2, ψ and β may be estimated directly via the function glm() of R and S. See Pregibon (1980) and McCullagh and Nelder (1989). The pertinent material is in McCullagh and Nelder (1989) on page 170. One sees there a multinomial probability mass function being represented as the product of binomials. One follows that representation in setting up the response and explanatory matrices for glm.

Now the basic probabilities are parameterized as

$$Prob\{Y = 2\} \ = \ 1 \ - \ exp\{-e^{\eta - \theta_2}\}$$

$$Prob\{Y = 1\} \ = \ exp\{-e^{\eta - \theta_2}\} \ - \ exp\{-e^{\eta - \theta_1}\}$$

$$Prob\{Y = 0\} \ = \ exp\{-e^{\eta - \theta_1}\}$$

This fits in with (3) via the connection

$$e^{-\psi} \ = \ e^{-\theta_1} \ - \ e^{-\theta_2}.$$

4.3. *The setup*

Suppose that:

$$\beta_i \text{ is the "strength" of team } i \text{ when playing at home}$$

and

$$\gamma_i \text{ is the "weakness" of team } i \text{ when playing away}$$

These will be assumed constant.

Now consider the model

$$Prob\{i \text{ wins at home playing } j\} \;=\; 1 \;-\; exp\{-e^{\beta_i+\gamma_j-\theta_2}\}$$

and

$$Prob\{i \text{ loses at home playing } j\} \;=\; exp\{-e^{\beta_i+\gamma_j-\theta_1}\} \qquad (4)$$

with the probability of a tie 1 minus the sum of these two.

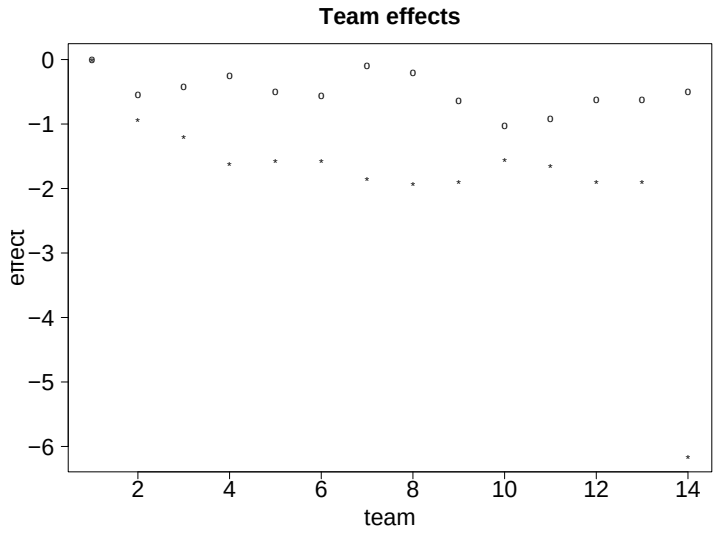

Figure 2 The estimated home effects, $\hat{\beta}_i$, are denoted "o" and the away, $\hat{\gamma}_i$, are "*".

In the fitting the results of the individual games will be assumed statistically independent. The fixed effects are meant to handle the connections amongst teams.

5. Results

The parametrization employed is (3). The estimation method is maximum likelihood. Figure 2 shows the resulting $\hat{\beta}_i$ and $\hat{\gamma}_j$ of (4). The values have been anchored by setting $\hat{\beta}_1,\ \hat{\gamma}_1 = 0$. One sees $\hat{\gamma}_{14}$ sitting near -6.15 in an attempt to get to $-\infty$ following losing all its away games. The residual deviance is 334.8 with $182 - 26 - 2 = 154$ degrees of freedom. The degrees of freedom here and later in the paper are those as if the model were fitted directly, i.e. the Pregibon trick was not employed. The away performances values stand out.

Table 4 Fitted values. The left hand columns refer to home games and the right hand to away. The fitted values are the number of games, 13, times the fitted probability

Team	W	T	L	W	T	L
1	8.35	3.10	1.55	10.10	1.52	1.38
2	5.96	3.26	3.78	6.73	2.95	3.32
3	6.40	3.25	3.36	5.56	3.31	4.13
4	6.91	3.20	2.89	3.73	3.65	5.62
5	5.96	3.21	3.83	3.87	3.63	5.50
6	5.76	3.19	4.06	3.83	3.63	5.54
7	7.44	3.13	2.43	2.79	3.67	6.55
8	6.96	3.21	2.83	2.44	3.61	6.95
9	5.41	3.14	4.45	2.45	3.60	6.95
10(a)	4.31	2.78	5.91	3.74	3.67	5.59
11(b)	4.60	2.90	5.51	3.36	3.68	5.96
12(c)	5.44	3.14	4.41	2.47	3.60	6.92
13(d)	5.42	3.14	4.44	2.47	3.61	6.92
14(e)	5.41	3.50	4.09	0.00	0.00	13.00

Table 4 gives the fitted wins-ties-losses for home and away. These number. Figure 3 plots fitted versus actual. One notes that the fitting definitely picks up Bryne losing all its away games. One also sees a clustering about the diagonal line in Figure 3.

6. Assessing fit

There is an issue of how to assess the fit of an ordinal response model. The link function may be checked by nonparametric regression, see Figure 11 in Brillinger et al (1980). Figure 4 shows the kernel estimate based on the data $(\hat{\eta}_i, y_i)$ where $\hat{\eta}_i$ is the fitted linear predictor and y_i is the observed Bernouli value. The smooth curve is the extreme value cumulative

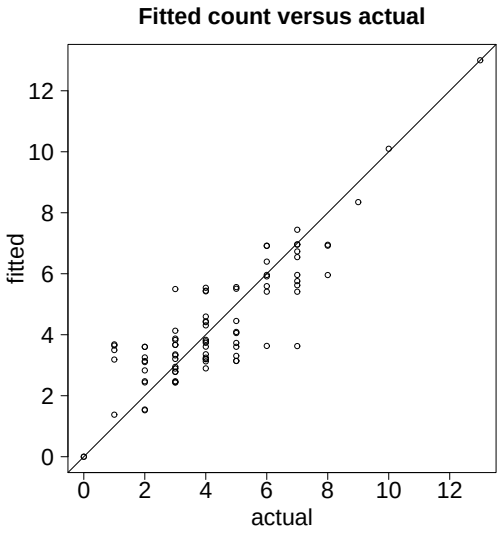

Figure 3 Fitted counts of wins, ties, losses against corresponding actual.

distribution function. The two follow each other.

6.1. *Chi-squared statistics*

It was indicated that the residual deviance of model (4) was 334.8 with 154 degrees of freedom, but the interpretation must be made with care. Further, one cannot simply interpret a chi-squared statistic based on the values of Tables 3 and 4 because the entries are negatively correlated following the competitive character of the variates - one wins, another loses.

6.2. *Uniform residuals*

> "The idea is to obtain randomized rank-sum statistics for the independence, randomness, k-sample and two factor problems analogous to the statistics of ... others." "... one essentially replaces the original data ... by a random sample ... known to have distribution ... advantage ... of having a continuous distribution ..."
>
> Bell and Doksum (1965)

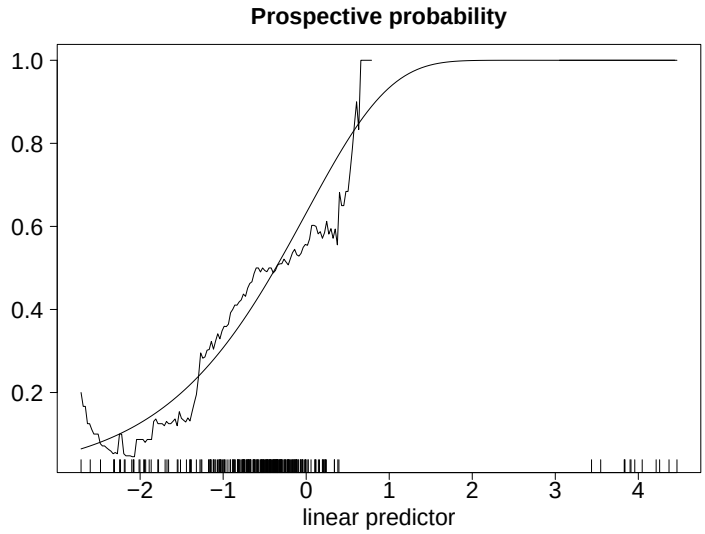

Figure 4 Prospective probabilities. The smooth curve is expression (2). There is a rug plot giving the linear predictor values.

In the case of a continuous variate, Y, the random variable $F(Y)$ has a uniform distribution, see Fisher (1932). Supposing the distribution depends on an unknown parameter θ with estimate $\hat{\theta}$, the $\hat{U} = F(Y|\hat{\theta})$ may be anticipated to have an approximate uniform distribution. The variates $\hat{U}_i = F(Y_i|\hat{\theta})$ were employed in Brillinger and Preisler (1983) to examine the overall fit of a model. They are an aid in various nonstandard cases, such as for random effect models.

In the present case the response employed is binary, $Y = 0, 1$ so various of the classical model effect procedures appear not particularly effective. In this binary case uniform residuals may be computed as follows.

Suppose

$$Prob\{Y = 1|explanatories\} = \pi$$

and that U_1 and U_2 denote independent uniforms on the intervals $(0, 1-\pi)$, $(1-\pi, 1)$, respectively . Then the variate

$$U = U_1(1-Y) + U_2 Y \tag{5}$$

has a uniform distribution on the interval $(0, 1)$. An effect of constructing these values is that the data that are 1's will become spread out in the upper interval and those that were 0's in the lower.

In the null case $E\{U\} = 1/2$ whereas when

$$Prob\{Y = 1|explanatories\} = \pi_0$$

then $E\{U\} = (1 + \pi - \pi_0)/2$.

In practice one has $\hat{\pi}$ an estimate of π and forms

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

where $\hat{U}_1$ and $\hat{U}_2$ are uniform on $(0, 1-\hat{\pi})$ and $(1-\hat{\pi}, 1)$ respectively. When an estimate of $\hat{\pi}$ is employed, we refer to $\hat{U}$, as a uniform residual.

One can equally employ normal residuals, $\Phi^{-1}(\hat{U}_i)$. Working with these has the advantage of spreading the values out in a familiar manner in the null case. We refer to $\Phi^{-1}(\hat{U})$ as a normal residual. Doksum (1966) uses the term "normal deviate" in the situation referred to at the beginning of this section. Various traditional residual plots may now be constructed using the $\hat{U}$ or $\Phi^{-1}(\hat{U})$, e.g. normal probability plots involving the normal residual, $\Phi^{-1}(\hat{U})$ versus an appropriate normal quantile and of $\Phi^{-1}(\hat{U})$ versus explanatories.

Discrete response cases were considered in Brillinger (1996) and Dunn and Smyth (1996).

6.2.1. *Results*

The normal residuals, $\Phi(\hat{U})$ were computed fitting the model (4) as discussed in Section 4. The idea is that they should have an approximate $N(0, 1)$ distribution if the model is fitting well.

The plots are shown in Figures 5 and 6. There are some indications of asymmetry. Next consideration turns to seeking to improve the fit by including possible explanatory variables. Available variables include: day of year, distance between cities, and results of the preceding game for the teams. Figure 7 provides a plot of normal residuals vs. day in year. A loess line, see Cleveland, Grosse and Shyu (1992) has been added. There is an indication of departure in the earlier part of the season. Figure 8 seeks to confirm this. It plots weekly totals of home wins by week of the season. There is a trend downwards as the season progresses. Figure 9 plots the residuals against the distance between the towns involved in the game. There is an indication of a bowing upwards.

An analysis of deviance was carried out fitting for the pair of teams in the game the explanatories of: the home team, the away team, day of game, distance between the towns involved and the results (W, T or L) of the teams' preceding game. The results are presented in Table 5. One sees that the visiting team and their previous week's result appear most important. The final deviance is 297.99 on 182-32-2 = 148 degrees of freedom. The degrees of freedom are less than in the previous case

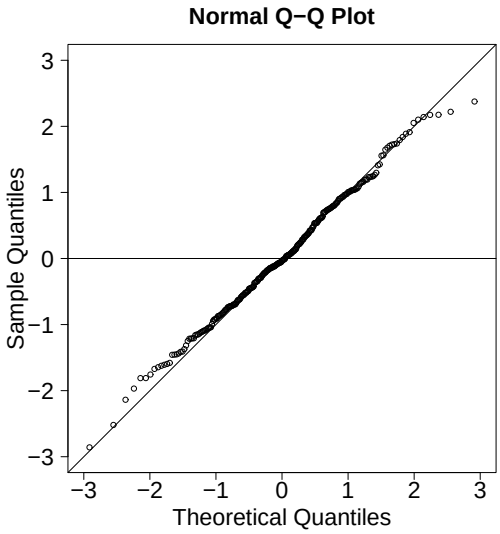

Figure 5 A normal probability plot of the "normal residuals".

because, in order to include the previous week's result, a week of data must be dropped.

Table 5 Analysis of deviance table for the inclusion of explanatories successively

Term	Df	Deviance change
Home team	13	8.76
Away team	13	42.59
Previous home	2	1.77
Previous away	2	4.26
Day	1	2.33
Distance	1	1.00

7. Another model

A simpler model is next considered. Let δ_i denote the strength of team i whether playing at home or away, i.e. assume $\beta = \gamma$.

In the computations expressions (4) are replaced by

$$Prob\{i \text{ wins at home playing } j\} = 1 - exp\{-e^{\delta_i - \delta_j - \theta_2}\}$$

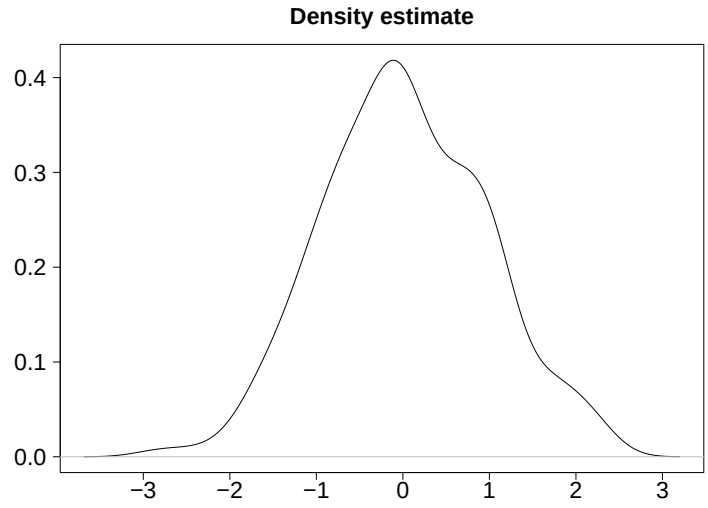

Figure 6 A kernel density estimate of the residuals.

and

$$Prob\{i \text{ loses at home playing } j\} \;=\; exp\{-e^{\delta_i - \delta_j - \theta_1}\}. \qquad (6)$$

The fitted values, $\hat{\delta}_i$, are given in Figure 7..

One sees in the figure that $\hat{\delta}_1$, the champion's strength, is particularly large while $\hat{\delta}_{14}$, the lowest team's is particularly small. This result is consistent with Figure 4.3..

The residual deviance of the model (6) is 357.36 with 182-14-2 = 166 degrees of freedom to be compared with the previous 334.8 with 154 degrees of freedom.

8. Uses

The fitted models obtained may be put to some uses. For example one could run Monte Carlos to estimate the probability of each team being champion or of being relegated, as in Lee (1997). Alternately one could examine the effects of a switch from 2 to 3 points for a win, again via simulation, if every thing else remained fixed.

Further, one could use the fitted models to assess various betting strategies. In that connection one referee mentioned that one could fit the model to say the first 20 week's data and then see how well that model predicts

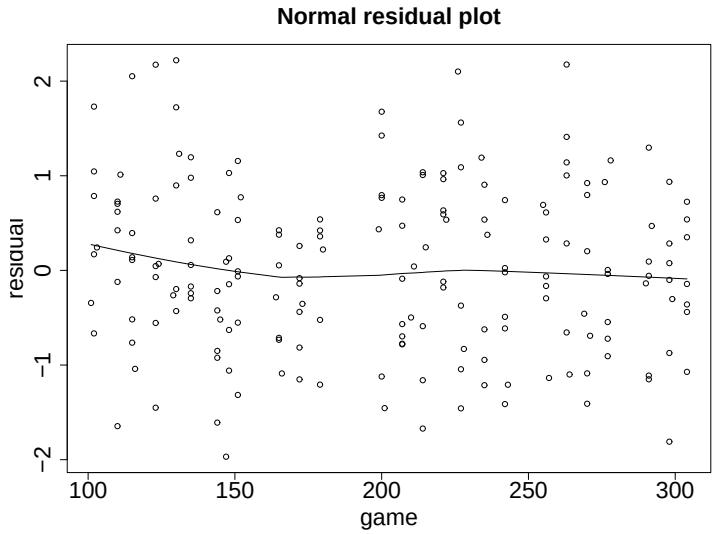

Figure 7 Normal residuals versus day of game for the home wins. A loess line has been added.

the next week's results. The other referee brought up the idea of using the model to rank the teams, but backed away because the design was completely balanced. I can add that if some explanatory with teeth could be found to include in the model, then ranking could proceed. The referee also added that if desired one could fit a single home advantage parameter for all the teams.

9. Extensions

A study was made of the effect of handling omitted variables by including additive random effects in the linear predictor. The resulting model corresponds to a different link function, for example the inverse link function becomes

$$1 - \int exp\{-e^{\eta+\sigma z}\}\phi(z)dz$$

instead of (2) if the effects are assumed to be $IN(0, \sigma^2)$. The resulting $\hat{\sigma}$ turned out to be near 0.

Harville (2003) and Stern (2004) are recent papers concerned with related problems for other sports and might be consulted by the interested

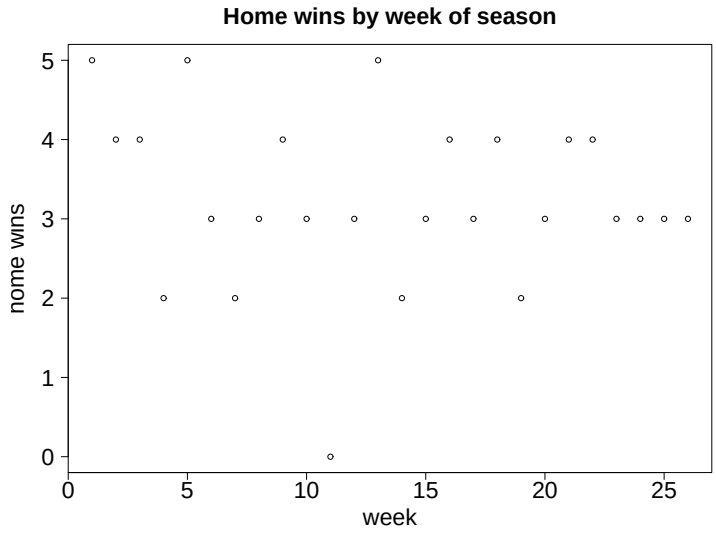

Figure 8 Actual number of wins versus week into season. There are typically 7 games a week.

reader.

10. Discussion and summary

Conditioning was employed to take the ordinal-valued case to a pair of conditionally independent cases. The advantage was that standard statistical packages became available for the analyses.

One could have modeled the goals and then obtained W-T-L results afterwards but the choice was made to try something different.

The fine away performance of Rosenborg and poor away performance of Bryne are perhaps the most notable features noted.

Acknowledgement. I thank Phil Spector and Ryan Lovett for help with the computing. I thank the two referees for important comments that improved the paper. The research was supported in part by NSF Grants DMS-02-03921 and DMS-05-04162.

Kjell, jeg snakker svaert lite norsk, men takk venn.

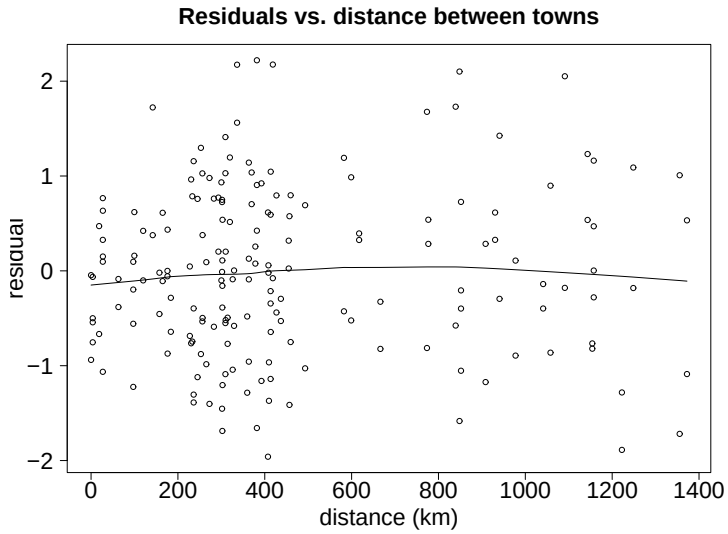

Figure 9 Normal residuals versus distance between towns. A loess line has been added.

References

1. Andersen, J. A. (1984). Regression of ordered categorical values. *J. Royal Statist. Soc.* B 46, 19-35.
2. Bell, C. B. and Doksum, K. A. (1965). Some new distribution-free statistics. *Ann. Math. Statist.* 36, 203-214.
3. Brillinger, D. R., Udias, A. and Bolt, B. A. (1980). A probability model for regional focal mechanism solutions. *Bull. Seismol. Soc. Amer.* 70, 149-170.
4. Brillinger, D. R. (1996). An analysis of an ordinal-valued time series, pp. 73-87 in *Athens Conf. on Applied Probability and Series Analysis. Volume II: Time Series Analysis.* Lecture Notes in Statistics, vol. 115. Springer-Verlag, New York.
5. Brillinger, D. R. (2005). Some data analyses using mutual information. *Brazilian J. Prob. and Statist.* 18, 163-183.
6. Brillinger, D. R. and Preisler, H. K. (1983). Maximum likelihood estimation in a latent variable problem. Pp. 31-65 in *Studies in Econometrics, Time Series and Multivariate Statistics.* (Eds. S. Karlin, T. Amemiya and L.A. Goodman). Academic Press, New York.
7. Cleveland, W.S., Grosse, E. and Shyu, W.M. (1992). Local regression models. Pp. 309-376 in *Statistical Models in S.* Eds. J. M. Chambers and T. J. Hastie. Wadsworth, Pacific Grove.

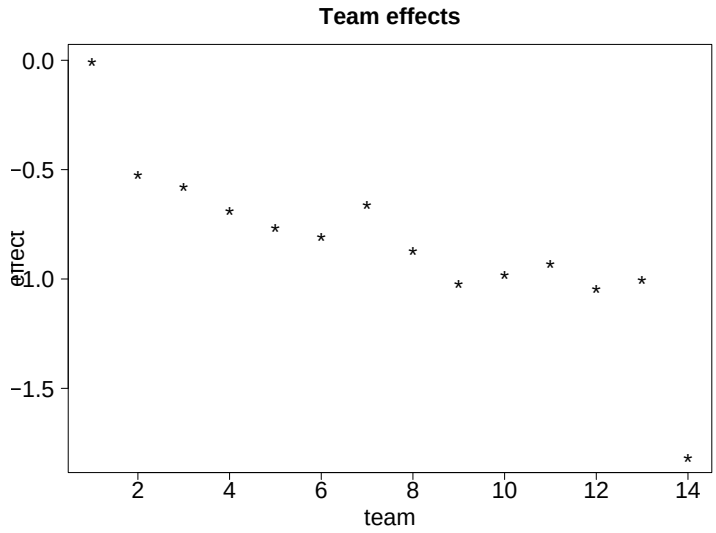

Figure 10 The $\hat{\delta}_i$, of the model (6), are indicated by "*"

8. Doksum, K. A. (1966). Distribution-free statistics based on normal deviates in analysis of variance. *Rev. Inter. Statist. Inst.* 34, 376-388.

9. Doksum, K. A. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Ann. of Statistics* 2, 267- 277.

10. Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *J. Computational and Graphical Statistics* 5, 236-244.

11. Dyte, D. and Clarke, S. R. (2000) A ratings based Poisson model for World Cup soccer simulation. *J. Operational Research Soc.* 51, 993-998.

12. Fienberg, S. E. (1980). *The Analysis of Cross-classified Data.* MIT Press, Cambridge.

13. Harville, D. (2003). The selection or seeding of college basketball or football teams for postseason competition. J. American Statistical Assoc. 98, 17-27.

14. Karlis D. and Ntzoufras J. (2000). On modelling soccer data. *Student* 3, 229-245.

15. Karlis, D. and Ntzoufras, J. (2003a) Analysis of sports data Using bivariate Poisson models *The Statistician* 52, 381-393.

16. Karlis, D. and Ntzoufras, J. (2003b) Bayesian and non-Bayesian analysis of soccer data using bivariate Poisson regression models. *16th Panhelenic Conference in Statistics.* Kavala, April 2003.

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

18. McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models.* Chapman and Hall, London.

19. Panaretos, V. (2002). A statistical analysis of the European soccer champion league. *Proc. Joint Statistics Meeting.*

20. Pearson, E. S. (1950). On questions raised by the combination of tests based on discontinuous distributions. *Biometrika* 37, 383-398.

21. Pregibon, D. (1980). Discussion of paper by P. McCullagh. *J. Royal Statist. Soc. B* 42, 139.

22. Stern, H. (2004). Statistics and the college football championship. *The American Statistician* 58, 179-185.